

City University London

MSc in Business System Analysis and Design

Project Report

2011

**Customer Churn prediction for an Automotive Dealership  
using computational Data Mining**

Vincenzo Selvaggio

Supervised by: Cristina Gacek

2012-01-06

## **Declaration**

*By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.*

*Signed:*

## **Abstract**

Customer retention has been extensively studied and implemented in the mobile industry based on the evidence that holding existing customers is less expensive than acquiring new customers. In fact, the strategic focus in any mature market needs to move from acquisition to retention. This is no exception for the automotive industry which is the focus of this project. The aim of the work is twofold: on one hand define the explanatory variables that forms the inputs of a data mining predictive model, on the other hand establish the best predictive model by comparing different mining algorithms.

For the case under study (an automotive dealership), 65 explanatory variables has been identified by applying a generic framework on customer loyalty, defined as a combination of customer behaviour and attitude. Those variables have been used as inputs for the machine learning techniques.

What's more, the ROC curve analysis reported that the best model is the Logistic Regression (area = 0.783), followed by Decision Tree (area = 0.638) and Neural Network (0.6148).

Finally, the results has suggested the automotive dealership the short and long term actions that should be taken in order to retain customers. The short term actions consist in creating new customer categories in order to have a diversification strategy to support individual needs. The long term actions are more time consuming and require changes in the current business processes.

## **Keywords**

Business Intelligence, Data Mining, Customer Churn, Automotive Industry, Classification Models.

## Table of Contents

<b>1</b>	<b>Introduction, Aim and Objectives .....</b>	<b>1</b>
1.1	Research Strategy.....	1
1.2	Project Structure .....	2
<b>2</b>	<b>Literature Review .....</b>	<b>4</b>
2.1	Knowledge discovery models .....	4
2.2	Classification methods .....	5
2.2.1	Decision Tree.....	7
2.2.2	Logistic Regression.....	9
2.2.3	Neural Networks.....	11
2.3	Customer Life Cycle .....	12
<b>3</b>	<b>Methodology.....</b>	<b>15</b>
3.1	Cross-Industry Standard Process for Data Mining.....	15
<b>4</b>	<b>Customer Retention in Automotive industry.....</b>	<b>17</b>
4.1	Business Understanding.....	17
4.2	Data Understanding .....	22
4.3	Data Preparation.....	31
4.4	Modelling.....	33
4.5	Evaluation and Deployment .....	43
<b>5</b>	<b>Conclusions.....</b>	<b>46</b>
5.1	Contribution and Limitation .....	46
5.2	Next Step.....	46
5.3	Data Mining and Ethics.....	47
	<b>References.....</b>	<b>48</b>
	<b>Appendix A.....</b>	<b>A1</b>
	Project Definition .....	A1
	<b>Appendix B.....</b>	<b>B1</b>
	Customer explanatory and target attributes.....	B1
	<b>Appendix C.....</b>	<b>C1</b>
	ORDER table .....	C1
	TRANSACTION table .....	C1
	CUSTOMER table .....	C1

CUSTOMER RULE table.....	C2
ITEM table .....	C2
ITEM ALTERNATIVE table .....	C2
<b>Appendix D .....</b>	<b>D1</b>
Interview Introduction.....	D1
Interview Question #1 .....	D1
Interview Question #1 - Notes .....	D2
Interview Question #2 .....	D3
Interview Question #2 - Notes .....	D3
Interview Question #3 .....	D4
Interview Question #3 - Notes.....	D4
Interview Question #4 .....	D5
Interview Question #4 - Notes.....	D5
Interview Question #5 .....	D6
Interview Question #5 - Notes.....	D7
<b>Appendix E.....</b>	<b>E1</b>
CD Content.....	E1
<b>Appendix F .....</b>	<b>F1</b>
Data Preparation SQL queries.....	F1

# 1 Introduction, Aim and Objectives

Any business organization keeps large amounts of data stored in their IT systems which is often referred as raw data as it belongs to several departments, it is stored in different formats and most likely redundant. When the data is consolidated in a centralized database (data warehouse) it becomes informational data. The activity that analyzes the informational data and creates additional knowledge for decision making is called Business Intelligence (BI).

BI may be implemented in two forms: passive and active. The passive form is carried out by querying the consolidated data (informational data), applying numeric techniques and showing visually the results. This form requires as such initial hypothesis and statistics methods to confirm it (or not). The active form instead requires the application of mathematical models to the informational data that generates additional knowledge not known a priori. Both forms may be used within the same decision-making system: the passive form identifies the variables later applied as the base for the mathematical models of the active form.

Data mining is one of the technologies to achieve Business Intelligence and it is applied in many contexts with a common goal: find patterns in existing data in order to get informational advantage for near future business and strategic decision.

This project focuses the attention on the context of customer relationship management (CRM) in the automotive industry. Customers are a key asset of any company and holding good customers, by providing incentive, is essential. BI helps in identifying which are good customers and most importantly predict which of those are going to leave for other competitors.

## 1.1 Research Strategy

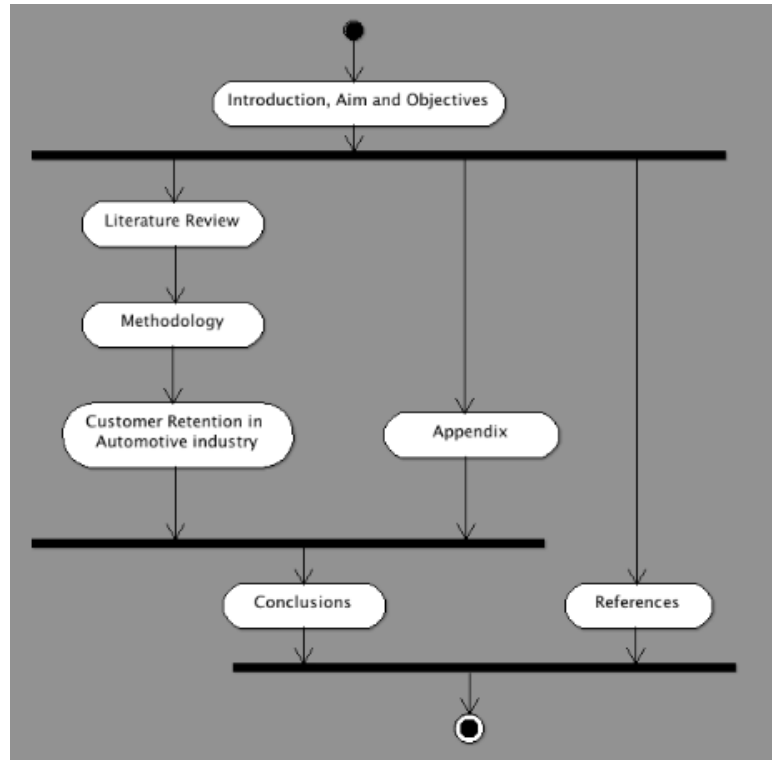
The project type (research strategy) is a combination of case study (client-based project) and academic research using quantitative data analysis. The aim of the project is to identify the variables that characterize the customers in the automotive industry and compare the different data mining models for predicting customer churn (customers that are likely to leave).

The beneficiaries of the project are the Client (the automotive dealership of the case study) and the Data Mining field. On one hand, the dissertation, by using the Client database as case study, provides a summary of the outcomes to be applied in business decision making. On the other hand, the data mining models comparison provides a contribution to the field by proposing which is the better predictor for customers churning in the automotive industry (generalization).

Therefore, the objectives can be summarized as follows: to conduct an analysis of the Client customer, product and transactional data; to deliver a summary of the findings to support the Client retention plan; to identify, build and evaluate classification data mining models for predicting customer churn in the automotive industry; to identify which model is a better predictor in the Client case study and attempt to generalize the theory to the automotive industry.

## 1.2 Project Structure

The project structure is shown as an Activity diagram (**Figure 1.1**), whereas each activity is a chapter of the dissertation.



**Figure 1.1:** *project activity diagram*

Introduction, Aim and Objectives is this chapter and presents the reader the scope of the project with its aim, objectives, research strategy and structure.

The Literature Review introduces in mathematical terms the classification data mining algorithms chosen as models to predict customer churn. Using a purposive sampling dataset, the theory is verified with simple numeric examples. In addition, the customer retention is described as a stage of the wider customer life cycle and a framework is proposed to study the factors that influence customer loyalty, meant as a combination of customer behaviour and attitude.

The Methodology chapter discusses the stages of a data mining project accordingly to the Cross-Industry Standard Process for Data Mining model (CRISP-DM) and explains the reason why a standardized process is important in this field.

The Customer Retention in Automotive industry chapter is the core of the project and reports the tasks and the outcomes of each data mining stage: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

In the Conclusions, the contribution to the field is stated, the limitation of the research acknowledged and the next steps suggested.

As the project structure adopted is not a conventional structure of a typical dissertation, a mapping between the proposed structure and the recommended structure is reported to facilitate the reader (**Table 1.1**). The decision of using a different structure has been taken in order to comply with the CRISP-DM standard process for data mining projects.

<b>Recommended Chapter</b>	<b>Proposed Chapter</b>
Chapter 1 - Intro and Objectives	Chapter 1 - Intro, Aim and Objectives
Chapter 2 - Engagement with Academic Literature	Chapter 2 - Literature Review
Chapter 3 - Methods	Chapter 3 - Methodology Chapter 4 - Customer Retention in Automotive industry: section 4.1 to section 4.5
Chapter 4 - Results	Chapter 4 - Customer Retention in Automotive industry: section 4.2 Data Understanding, section 4.4 Modelling
Chapter 5 - Discussion	Chapter 4 - Customer Retention in Automotive industry: section 4.5 Evaluation and Deployment
Chapter 6 - Evaluation, Conclusions	Chapter 4 - Customer Retention in Automotive industry: section 4.4. Modelling, section 4.5 Evaluation Chapter 5 - Conclusions

**Table 1.1:** *project structures comparison*

Proposed Chapter 1 and 2 match the recommended ones.

Chapter 3 introduces the methodology to conduct the project whilst the actual methods are described in Chapter 4 (section 4.1 to 4.5).

The results are presented in Chapter 4, in particular they are the outputs of the two stages Data Understanding and Modelling (section 4.2, section 4.4).

The discussion of the results are reported in the Evaluation and Deployment stage (section 4.5 of Chapter 4); here is evident the comparison with the objectives set in the introduction.

The results are evaluated firstly from numeric point of view (section 4.4 Modelling) using the 10-fold cross validation technique and the ROC curve analysis, then from a business perspective (section 4.5 Evaluation) with an interview.



## 2 Literature Review

This chapter introduces the reader to the main concepts of a data mining model and covers the theory behind the classification techniques chosen as possible models to predict customer churn. Moreover customer retention is introduced as application of business intelligence and a framework is presented to identify the factors that affect the customer loyalty and serve as inputs for the mining algorithms. Data mining is a well established field and the concepts proposed may be found in many articles and books at different level of details. The main references utilized are Fayyad et al. (1996) for the introduction on discovery models and Vercellis (2009), Berry and Linoff (2004) for the details of the algorithms.

### 2.1 Knowledge discovery models

Fayyad et al. (1996, p82) define the knowledge discovery in databases (KDD) as *'the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data'*. Therefore, data mining or also called knowledge discovery is a set of activities aiming at analyzing large databases and extracting extra information meaningful for decision making or problem solving (Vercellis, 2009). As described in more details in the methodology chapter, among the several data mining tasks the core activity consists in developing an inductive learning model whereas the main purpose is to generalize patterns or rules stem from the samples available.

Those learning models are mathematical models and they represent a symbolic representation of real problems. They are divided in two types of goals: prediction and description. The former goal concerns in finding patterns with the objective of predicting future outcomes, the latter is to present the results of the discovery in human readable format.

Both goals, as stated by Fayyad et al. (1996) in their attempt towards a unifying framework for KDD, can be achieved with the following class of methods.

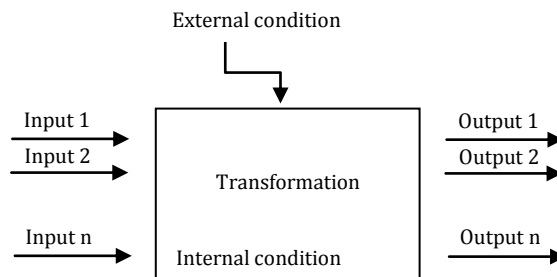
Classification: this is one of the most common knowledge discovery task that consists in establishing predefined categories, deriving a learning function that maps existing samples to one of the category (or called class) and assign newly presented sample to one of the class based on the derived learning function. An example of classification is the risk evaluation that a bank calculates for a new customer when asking for a loan. Based on previous applicants' behaviour, the bank classifies the new request as low, medium or high risk and therefore accepts or rejects the application.

Regression: similar to classification, a function is derived to map a sample to a continuous value (instead of discrete value or class). This method is also referred as Estimation task because of the real valued prediction variable that the function estimates, for example predict a customer electricity bill so that is as close as possible to the real usage.

Clustering: the task of identifying and segmenting the items into a finite number of categories which are not predefined (unlike classification). The items are grouped together based on their homogeneity and it is up to the data miner to give a meaning to the results. The Clustering task may be done as part of a Classification data mining problem, whereas the first step is to decide the predefined classes which in some cases may not be obvious to come up with (Berry and Linoff, 2004).

Association Rule: the task of determining recurring relationship between groups of samples, for this reason it is also called Affinity Grouping. The classical application of this model is finding out which products are purchased together by customers and propose the same combination of items to new customers. Amazon successfully used this technique to increase their sales but also giving appreciated suggestion to the users.

Regardless of the method implemented, a data mining learning model can be represented as a system (see **Figure 2.1**), in other words it receives a set of inputs, transform those inputs based on external and internal conditions and produce a set of outputs (Mallach, 2000). Understanding the inputs and the outputs of such a system is as important as knowing what happens in between.



**Figure 2.1:** *model of an input-output system*

The input takes the form of a two dimensional table and is named dataset. Each row of the table correspond to an instance (also referred as sample, observation, record or generally item) and each instance is characterized by the values of the attributes (the columns of the table).

There are two types of attribute: categorical and numerical (Stevens, 1946). On one hand, the categorical ones take on values in a discrete set of possibilities, for example a customer city of residence, and arithmetic operations are not allowed. On the other hand numeric attributes measure numbers either as real or as integer and arithmetic operations are permitted. For example, the number of transactions per week per customer.

If one of the attribute is the target variable, which means it is the attribute to be predicted for new instances but are known for existing instances, the learning process is called supervised. This is the case for Classification and Regression models but not for Clustering schemes which instead are called unsupervised and are not guided by a target attribute. In the latter case, as explained above, the purpose is to identify clusters of instances that have similarities.

Unlike the input, the output may have several forms and it depends on the method chosen. The output of a linear regression model is just a number that is the weighted sum of the input attributes; a decision tree may be a graphical representation of a classification model; a set of rules in the form *if a and b then x* is the output of an association rule algorithm (Branchman and Levesque, 1985).

## 2.2 Classification methods

The most widely used example of classification problem in literature is the customer churn in the mobile industry, which is introduced as the basis for explaining how the algorithms work.

To keep the case simple, the input of the model is represented by the **Table 2.1** whereas there are 4 explanatory attributes and 1 target attribute (the class Loyalty).

Area	Age	Incoming calls	Outgoing calls	Loyalty
A	<30	24	55	0
B	<30	222	343	0
A	>30	43	66	1
C	>30	32	88	1
B	>30	150	34	0
B	>30	98	76	1

**Table 2.1:** *purposive sampling dataset*

Area and Age are demographic and personal information whilst Incoming and Outgoing calls relates to the use of the service provided. As said above, those are called explanatory since they characterize each instance in the table. The last attribute Loyalty is the binary class which says that the customer is a churner or not based on a value calculated in a subsequent period. To better understand the difference of periods in the calculation of the two type of variables (explanatory versus target), say that the explanatory variables were calculate for the period May-June and the loyalty calculated in August. In May-June that specific customer (a row in the table) was still active considering the number of calls but then marked as disloyal if for example the number of total calls dropped below a threshold, for instance less than 10, in August. The data mining task is to establish a relationship (a function) between the inputs (the first 4 columns) and the output (the binary class) based on the existing samples and apply the function to classify future observations.

In mathematical terms, given a table of  $m$  instances and  $n$  attributes, we can represent the input dataset  $D$  as matrix:

$$X = [x_{ij}], i \in M, j \in N$$

whereas

$$\begin{aligned} x_i &= [x_{i1} \ x_{i2} \ \cdots \ x_{in}] \\ a_j &= [a_{1j} \ a_{2j} \ \cdots \ a_{mj}] \end{aligned}$$

are respectively the  $i$ -th row vector and the  $j$ -th column vector.

The vector  $x = [x_{i1} \ x_{i2} \ \cdots \ x_{in-1}]$  represents the so called predictive or explanatory values and  $y = x_{in} \in H = \{v_1, v_2, \dots, v_H\}$  is the target that may assume  $H$  distinct values.

A classification problem is to identify a function  $f(x)$  so that  $f(x): R^{n-1} \rightarrow H$  with the best accuracy (Vercellis, 2009).

To calculate the accuracy, the dataset  $D$  is split in two dataset  $T$  and  $V$  with  $D = T \cup V$ , respectively called training set and valuation set.  $T$  dataset is used to determine the functions  $f(x)$  and  $V$  to evaluate which of those have better accuracy defined as:

$$acc(V) = 1 - \frac{1}{v} \sum_{i=1}^v L(y, f(x))$$

with  $v$  number of instances in the dataset  $V$  and

$$L(y, f(x)) = 0 \text{ if } y = f(x) \quad L(y, f(x)) = 1 \text{ if } y \neq f(x)$$

### 2.2.1 Decision Tree

Decision trees are popular in the data mining field due to the fact that a tree is ultimately a set of rules and are easily interpreted or expressed in English as opposed to other techniques where the results cannot be explained (like neural networks).

The rules split the large dataset into smaller homogeneous groups with the respect to a particular target class therefore this algorithm takes also the name of divide-and-conquer partitioning scheme (Berry and Linoff, 2004). To understand how a record is classified using a tree, the rules can be thought as a set of questions: at the root node of the tree the first question is asked to decide the child node to enter; once in the child node the next question is asked to decide which child of the child node to enter and so on; once arrived at a leaf of the tree the record is associated with the class for that leaf. Different leaves can have the same class but for different reasons since the questions were answered differently along the path from the root to the leaf.

The construction of the tree is based on a recursive algorithm (Quinlan, 1986) that splits the dataset at each node (including the root node with the starting dataset  $T$ ) into smaller dataset applying a splitting criterion. A good split is the one that has high purity which means that the datasets generated contain instances with a predominant target class.

For example, let's consider the **Table 2.1** proposed above and in particular the two variables Area and Age. The first step of the algorithm is to define which of the two is a better first split with regards to the target class Loyalty. At high level, the Age splits the initial dataset  $T$  in 2 subsets while Area in 3 subsets as show in **Table 2.2** and **Table 2.3**. The Age is a better split: each of the descendant dataset has a predominant target class.

Input = Age	Total	Target class 0 (disloyal)	Target class 1 (loyal)
> 30	4	1	3
< 30	2	2	0

**Table 2.2:** dataset split using the variable Age

Input = Area	Total	Target class 0 (disloyal)	Target class 1 (loyal)
A	2	1	1
B	3	2	1
C	1	0	1

**Table 2.3:** dataset split using the variable Area

To formalize the decision tree algorithm, say  $X_j$  the  $j$ -th explanatory variable. If  $X_j$  is a categorical attribute the instances are categorized in  $K$  disjoint subsets with  $K$  the number of distinct values of the attribute. In the case above, for the variable Age we have  $K = 2$  (the 2 values  $>30$  and  $<30$ ) therefore 2 subsets  $B_0$  and  $B_1$ .

Among the  $X_j$  available attributes, the choice of the best split is based on an evaluation function or in other words the purity measure conceptually introduced above. There are different evaluation functions all trying to achieve the same effect, for the sake of the argument the *Gini* index is proposed since it is one of the most used.

Let's consider  $p_h$  the proportion of instances in the dataset  $T$  corresponding to the value  $v_h$  of the target class, the *Gini* index is defined as:

$$Gini(T) = 1 - \sum_{h=1}^H p_h$$

This is the purity measure of the parent node and it is compared to the purity measure derived from the child nodes for the given explanatory attribute. If the explanatory attribute selected divided the dataset  $T$  in  $B_k$  subsets, the *Gini* index of the child nodes is defined as:

$$Gini(B_0, B_1, \dots, B_k) = \sum_{k=1}^K p_k Gini(B_k)$$

whereas  $p_k$  is the proportion of instances in the subset  $B_k$  compared to the parent dataset  $T$ .

The explanatory attribute chosen among the one tested is the one that maximize the difference

$$\Delta = Gini(T) - Gini(B_0, B_1, \dots, B_k)$$

in other words higher purity.

Considering the example above we have for the root node

$$Gini(T) = 1 - (3/6)^2 - (3/6)^2 = 0.50 \text{ (lowest purity)}$$

for the explanatory variable Age

$$Gini(B_0) = 1 - (1/4)^2 - (3/4)^2 = 0.38$$

$$Gini(B_1) = 1 - (2/2)^2 - (0/2)^2 = 0.00 \text{ (highest purity)}$$

$$Gini(Age) = Gini(B_0, B_1) = 4/6 * 0.38 + 2/6 * 0.00 = 0.25$$

$$\Delta(Age) = Gini(T) - Gini(Age) = 0.25$$

for the explanatory variable Area

$$Gini(B_0) = 1 - (1/2)^2 - (1/2)^2 = 0.50$$

$$Gini(B_1) = 1 - (2/3)^2 - (1/3)^2 = 0.44$$

$$Gini(B_2) = 1 - (0/1)^2 - (1/1)^2 = 0.00$$

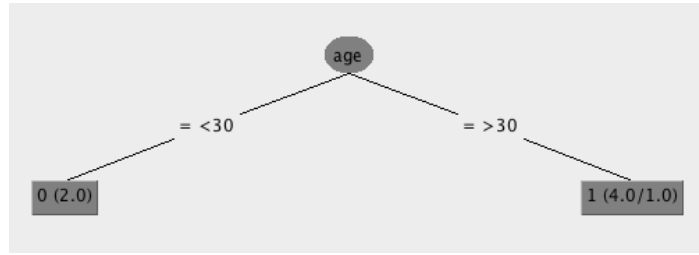
$$Gini(Area) = Gini(B_0, B_1, B_2) = 2/6 * 0.50 + 3/6 * 0.44 + 1/6 * 0.00 = 0.38$$

$$Delta(Area) = Gini(T) - Gini(Age) = 0.12$$

Comparing the two deltas, the Age is a better split since it maximizes the difference of purity.

The decision tree algorithm splits each node as long as there are enough elements to create other branches, however a tree with many branches loses generality and becomes more specific to the  $T$  dataset used (overfitting). For this reason an algorithm has a set of predefined rules (for example a purity threshold) that decides if the branching process should be stopped at any given node (this process is called pre-pruning). Another way to reduce the number of branches is to use a post-pruning approach, in practice, at the end of the construction of the tree, remove or merge nodes without undermining the accuracy of the model (Vercellis, 2009).

In **Figure 2.2** is shown the output of the decision tree generated by the software Weka using as input the dataset  $T$  (**Table 2.1**), considering only the attributes Age and Area.



**Figure 2.2:** decision tree graphical representation from Weka

As per our calculation above, the algorithm has chosen the variable Age as rule which correctly classifies the class 0 and only have 1 misclassification for the class 1 (4.0/1.0 means that of 4 instances in the leaf of class 1, 1 was incorrect).

Weka is a popular Java based machine learning suite adopted to perform the data mining tasks required by this project.

### 2.2.2 Logistic Regression

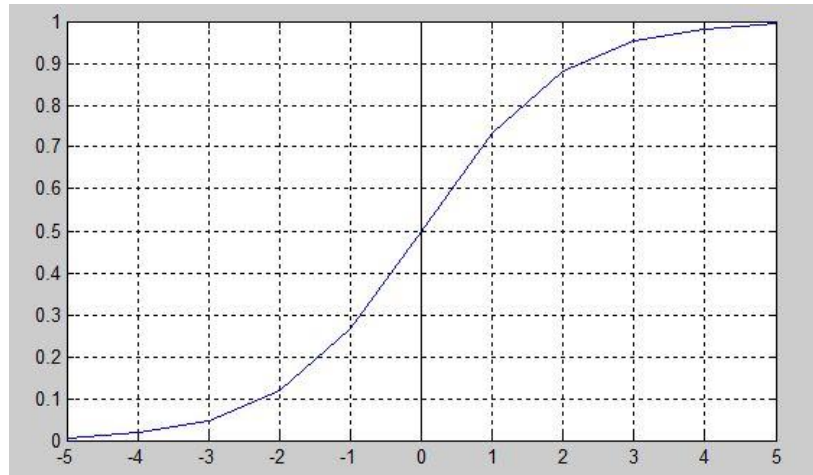
Logistic regression is a useful technique in cases the target class is a binary variable ( $H = \{v_0, v_1\}$ ) and the attributes are numeric (Rud, 2001). It is used to calculate the conditional probability of the target class given the set of explanatory variables by using the logistic function defined as:

$$p = f(wx) = Pr\{y = 0|x\} = \frac{1}{1 + e^{-wx}}$$

with

$$wx = w_0 + w_1x_1 + w_2x_2 + \dots + w_{n-1}x_{n-1}$$

where  $w_0$  is called intercept and  $w_1, w_2, \dots, w_{n-1}$  are the regression coefficients of the input variables  $x_1, x_2, \dots, x_{n-1}$ .



**Figure 2.3:** plot of a logistic function using Matlab

The logistic function, also referred as S-shaped curve, is plot in **Figure 2.3** using a numerical computing environment called Matlab. Intuitively, it is fair to say that for any value of the input  $wx$ , the output can only assume probabilities value from 0 to 1 (y axis).

Applying the transformation:

$$\ln\left(\frac{p}{1-p}\right) = wx = w_0 + w_1x_1 + w_2x_2 + \dots + w_{n-1}x_{n-1}$$

the calculation of the weights is converted to a linear regression model that can be easily resolved using the least squares approach.

Therefore the output of this model is a set of coefficients (vector  $w$ ) that are applied to the attributes of a new instance (vector  $x$ ) to obtain the target class to which it belongs to (0 or 1).

In **Figure 2.4** the output generated by Weka applying the logistic function (called SimpleLogistic) to the dataset  $T$  of **Table 2.1**, once more considering only Age and Area for simplicity.

```

Classifier output
=== Run information ===
Scheme:      weka.classifiers.functions.SimpleLogistic -I 0 -M 500 -H 50 -W 0.0
Relation:    QueryResult-weka.filters.unsupervised.attribute.Remove-R3-4
Instances:   6
Attributes:  3
              area
              age
              loyalty
Test mode:   evaluate on training data
=== Classifier model (full training set) ===
SimpleLogistic:
Class 0 :
1 +
[age] * -1.5
Class 1 :
-1 +
[age] * 1.5
Time taken to build model: 0.03 seconds

```

**Figure 2.4:** SimpleLogistic classifier output using Weka

Age is considered as variable  $x_1$  to use as input and the coefficients calculated are  $w_0 = 1$  and  $w_1 = -1.5$  for Class 0 (Age > 30) and  $w_0 = -1$  and  $w_1 = 1.5$  for Class 1 (Age < 30).

To validate the model, consider the first row of the dataset  $T$  having Age > 30.

The  $wx$  value is calculated by

$$wx = 1 + 0 * -1.5 = 1$$

consequently the logistic value is

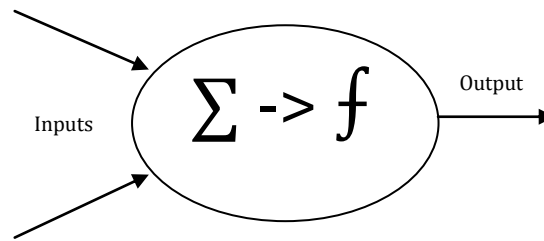
$$f(wx) = \frac{1}{1 + e^{-1}} = 0.7311$$

which maps the instance examined to the Class 1 by means of the probability of 73% of being true.

### 2.2.3 Neural Networks

Conceptually a neural network tries to emulate the behaviour of connected neurons each of one transforming multiple inputs in a single output.

The function that transforms the inputs in output takes the name of activation function and has two part: the combination function (usually a weighted sum of the inputs) and a transfer function that converts the single combined value to the output value (see **Figure 2.5**).



**Figure 2.5:** single neuron of a neural network

Accordingly to this definition the presented logistic regression can be modelled as a simple neural network (having just one neuron). Indeed one of the major factors that this method became popular in the data mining field is its close relation to probability and statistics (Berry and Linoff, 2004).

In mathematical terms, given the vector  $x$  of the explanatory attributes, the combination function is defined by

$$c(x) = w_1x_1 + w_2x_2 + \dots + w_{n-1}x_{n-1} + b$$

whereas  $b$  is an additional constant referred as bias or offset,

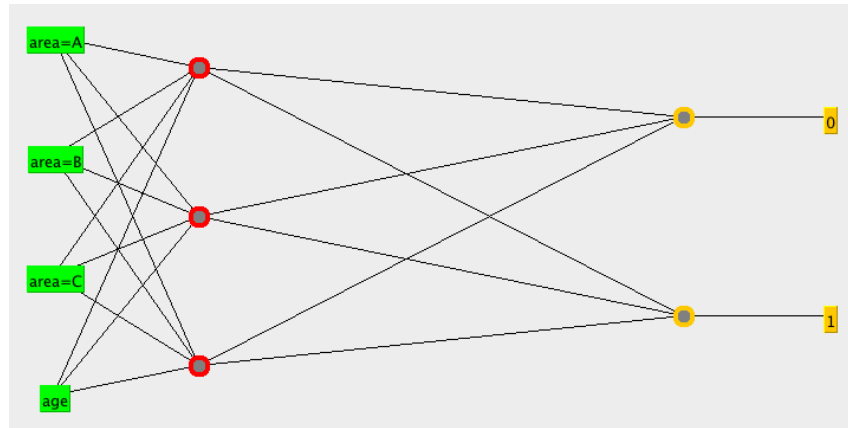
whilst the transformation function as

$$y = f(x) = t(w_1x_1 + w_2x_2 + \dots + w_{n-1}x_{n-1} + b)$$

For a neural network, intended as a collection of linked neurons presented in the formulae above, the learning process is done using the so defined back propagation which consists in the following steps: for each new instance of the  $T$  dataset, calculate the final output of the network, therefore derive the error between the calculated result and the expected one and finally adjust the weights of all neurons to minimize the error. Ultimately, given a combination function  $c(x)$ , a transformation function  $f(x)$  and a dataset  $T$ , the aim is to identify the vector  $w$  of the weights for all the neurons.



**Figure 2.6** represents the neural network generated in Weka using the multi-layer perceptron algorithm. This type of network takes its name for the hidden layer between the inputs and the output node: all inputs are connected to each hidden node and all hidden nodes converge to the output node (Bishop, 1995). The calculated coefficients produce the same predictions results of the logistic regression and the decision tree due the simplicity of the dataset  $T$ .

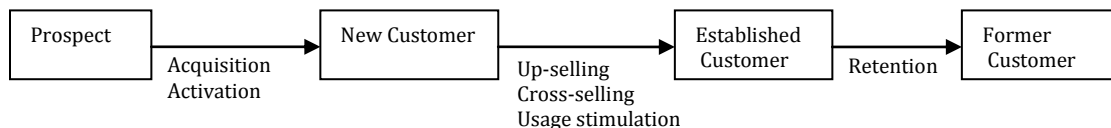


**Figure 2.6:** back propagation neural network generated by Weka

## 2.3 Customer Life Cycle

There are countless applications of data mining techniques involving a broad range of fields: web mining, credit evaluation, weather forecasting, electromechanical device diagnosis, biology, astronomy, just to mention a few. In the last decade one of the most active field is the area of customer relationship management (CRM) although the role of knowledge discovery can be considered complementary to the other existing and consolidated customer service initiative.

Customers are the most valuable asset for most business and it is important to understand their behaviours and to think of them as changeable entities (Berry and Linoff, 2004). Those changes are phases that represent the different relationships between the customer and the company and they take the name of customer life cycle. As shown in **Figure 2.7** there are 4 majors stages and several business processes involved.



**Figure 2.7:** customer life cycle

The *prospect* can be defined as a potential customer not yet acquired. Once a prospect make a first transaction with the company she becomes a *new customer* while a customer having more than one transaction can be considered an *established customer*. If the customer stops the relationship with the company his status changes to *former customer*.

A company models its business processes around those stages with the aim of moving the customer relation from one stage to the next.

Customer acquisition is the process of identifying the target market and the role of data mining is to identify the individuals with the higher probability of response. The obvious benefit is minimizing the costs and acquiring good customers.

Once a prospect shows interest, the next step is the activation, in other words some sort of subscription or payment needs to take place. A percentage of customers are lost in this phase and the knowledge discovery algorithms may help in understanding the reason why this has happened consequently improving the operational side for future acquisitions.

For established customers the business processes (called CRM) are aimed at maximizing the customer value, for example by selling premium services (up-selling), more products (cross-selling) or simply ensuring that the customers makes more transactions (usage stimulation). Those three activities are well supported by data mining predictive models that ensure the right actions are taken for that particular customer.

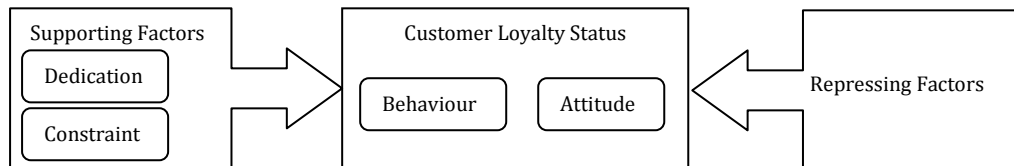
Predictive techniques are not only useful to understand the good customer and the way of improving the sales but most importantly to determine which customer is about to leave. The retention is a critical process due the fact that acquiring new customers is more expensive than keeping old customers not to mention that at the beginning of the relation new customers generate less value compared to established ones (Buckinx and Van Den Poel, 2004). Ahn et al. (2006) state that the strategic focus in mature market needs to move from acquisition to retention.

Hadden (2008) in reviewing the work of Bernet et al. (2001), Burez and Van Den Poel (2008) and Kim and Yoon (2004) describes the different types of churning customers. A first distinction is between non-voluntary and voluntary churn. The former occurs when the company terminates the service rather than the customer. The main reasons are abuse of the service or unpaid bills. The voluntary churners instead can be subcategorized in incidental and deliberate. Churners are considered incidental when the withdraw of the service is due to the fact the customers move to another city where the service is not available or their financials do not allow them to support that service anymore. Deliberate churners are customers that moved to a competitor and those are the ones to detect and try to retain and the central point of this project.

Nordman (2004) defines a framework to analyze the factors that influence customer loyalty and disloyalty. Those factors are divided in two macro categories, the loyalty-supporting factors which has a positive effect on the customer loyalty status, therefore strengthening it, and loyalty-repressing factors which decrease the customer loyalty status. The loyalty status is meant as a combination of customer behaviour and attitudes with the assumption that changes over time.

Loyalty-supporting factors can be further divided in two non exclusive subcategories: constraint and dedication loyalty (Bendapudi and Berry, 1997). The former is a type of relationship between the customer and the provider based on the fact that the customer does not perceive other better alternative, while the latter is a relationship based on the will of customer to stay loyal. Both are important: on one hand the constraints make the relationship persist and it is positive for the behaviour component of the loyalty status, on the other hand the dedication makes it grow therefore

positive for the attitude of the customer (Nordman, 2004). Loyalty-repressing factors are those that have a negative impact on both behaviour and attitude of the customer therefore one may think they are the direct opposite of the supporting factors. Different studies from the literature review conducted by Nordman (2004) suggest that based on empirical evidence the repressing factors do have their own categorization and they do not necessarily exist in a customer-supplier relationship. **Figure 2.8** shows the framework that is adopted later in the project to identify the customer data required by the data mining predictive algorithms.



**Figure 2.8:** *factors affecting the customer's behaviour and attitude*

### 3 Methodology

The need of a standard process for conducting a data mining project is twofold.

The first important reason is repeatability. Having an iterative process, consisting in a sequence of stages, help the data miner to conduct the same project more than once, applying it to a different set of data for the same company or to a different company by revisiting some of the early stages. As result, if on one side it makes the next projects faster to execute on the other side it also gives the miner confidence on the validity of future results.

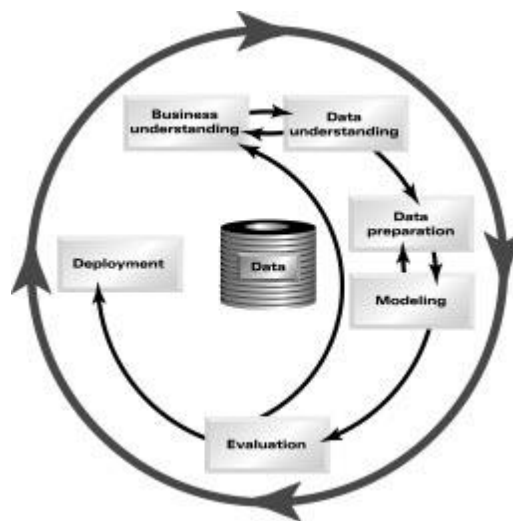
The second goal of such a process is to avoid two dangerous outcomes of the mining tasks: learning patterns that aren't true and learning patterns that are true but not useful (Berry and Linoff, 2004). For example, rules may be derived from a set of data, but this data is not at the right level of details or somehow truncated due to summarization, consequently not a good set of inputs. Other undesirable outcomes are finding relationships that are already known or relationships which were unknown but cannot be used in any business decision.

#### 3.1 Cross-Industry Standard Process for Data Mining

Many process models have been proposed in literature, each having different focus on certain aspects. For instance, in their unifying framework for KDD (knowledge discovery in databases), Fayyad et al. (1996) approves the practical model designed by Brachman and Anand (1996) with the emphasis on the interactive and iterative nature of it.

The methodology chosen for this project is the CRISP-DM (Cross-Industry Standard Process for Data Mining) model which accordingly to Shearer (2000) is a complete blueprint for conducting data mining projects, developed by consortium of industry leaders having an interest in creating a widely accepted approach. Strictly speaking, an industry neutral, tool neutral, well document model to enable miners to obtain better results and use best practices.

As shown in **Figure 3.1** the CRISP-DM consists in 6 phases: *Business understanding*, *Data understanding*, *Data preparation*, *Modelling*, *Evaluation* and *Deployment*.



**Figure 3.1:** CRISP-DM process model (Shearer, 2000, p14)

The outer circle around the model represents the cyclical aspect of a data mining project, in the sense that it never ends: once an initial sets of questions are answered other new questions arise. The internal arrows indicate the dependency among the different stages.

The first stage is *Business understanding* which consists in analyzing the requirements from a business point of view. Together with the client, the output of this phase is to delineate the main business goal of the project and turn it to a data mining problem with a set of explicit questions to answer.

Phase two is named *Data understanding* where the focus is familiarizing with the data. There are 4 tasks to perform: collect the available data which could have many sources, describe the data making sure the information is relevant in respect to the requirements, explore the data to form initial hypotheses answering the data mining questions delineated, finally verify that the quality of the data so that it does not produce wrong answers.

The *Data preparation* phase has the main purpose of producing the final dataset to feed to the modelling algorithms. As such, the data miner needs to select the necessary data based on the questions to answer and transform the data in the required format of the mining algorithms.

Transformation may consist, for example, in creating a derived table from multiple ones or creating a new field resulting from a combination of existing attributes.

The *Modelling* phase coincides with the application of different algorithms for the same data mining problem. Each model may require different types of data input, hence the *Data preparation* stage can be repeated. Once the models are built, the data miner must test the model's quality and the validity. Typically the dataset is divided in 2 sets, the training set to build the data and the test set to calculate the accuracy.

While in *Modelling* phase the validity of the model is evaluated from an accuracy point of view, in the *Evaluation* phase the data miner with the help of the business analyst verifies the results in the business context, especially if the initial questions are properly answered and if there is a way to make use of the findings in business decision. Moreover, all the stages are reviewed to certify that the results are not biased by incorrect assumption. At the end, the business analyst decides if another iteration is required to refine the answers or to proceed with the final stage.

The creation and validation of the model is not the end of the project. A final stage is required called *Deployment* having as outcome a report, presenting the knowledge acquired in such a way that the Client can understand and make use of it.

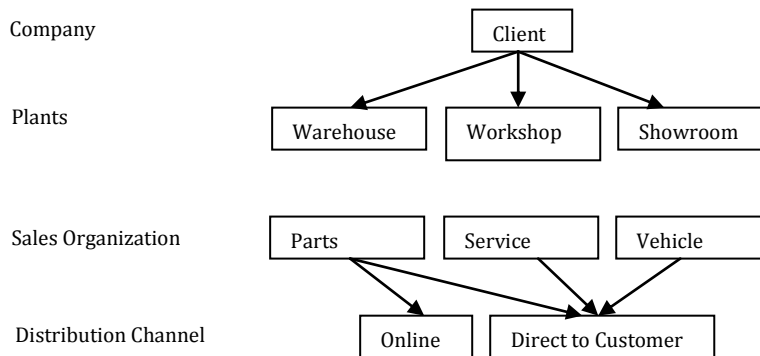
## 4 Customer Retention in Automotive industry

This chapter describes the activities and related outputs of the data mining stages applied to the presented case study. It is evident the linking between the practical tasks and the theory described in the literature review, in particular the customer loyalty framework and the classification learning algorithms respectively applied to the business understanding stage and, after the required data preparation, to the modelling phase. The chapter concludes producing and presenting the findings whose validity is opportunely verified on one hand by using the ROC curve accuracy indicator and on the other hand subjective business criteria.

### 4.1 Business Understanding

The first important output of this stage of the project is the gathering of the background information of the business under analysis.

As mentioned in the introduction, the Client is an automotive dealership whose business is focused on spare parts direct sale, service and sales of commercial vehicle (mainly trucks of all range). The parts and vehicles sold from Client to customers are sourced from a main supplier (which it is referenced in the project as Supplier I) and an alternative supplier (Supplier V producing not-OEM spare parts of Supplier I). The organization can be represented with the diagram in **Figure 4.1**.



Sales Organization	Distribution Channel	Plants		
		Warehouse	Workshop	Showroom
Parts	Online	x		
Parts	Direct to Customer	x		
Service	Direct to Customer	x	x	
Vehicle	Direct to Customer			x

**Figure 4.1:** Client organizational structure

The plants are the places where goods and services are provided therefore it is clear to see that the warehouse is the plant where the spare parts are kept in stock, the workshop is the plant where the service work (repairs of the vehicles) is carried out and the showroom whereas the trucks are parked.

The sales organizations are the departments responsible for the sales of the goods or services and there are 3 units: Parts, Service and Vehicle. As show in the diagram, the Parts sale unit can sell either directly to the customer or online, while Service and Vehicle only directly to the customer.

Finally the table shows how the sales organizations and distribution channels are related to the plants. Parts sale unit uses the warehouse plant, Service both the warehouse and the workshop since it needs spare parts to perform the service work and the Vehicle unit uses the showroom parking space.

The business is supported by an ERP (enterprise resource planning) system that integrates all the Client processes: inventory management, sales and distribution, finance, customer relationship management. The main advantage of having such a system is the information sharing across departments that eliminates redundancy and improves efficiency and productivity. From the data gathering point of view this facilitates the data miner task of consolidating the information since all information come from a centralized database and not from multiple sources.

From the business point of view (the Client), the primary goal of this data mining project is to identify the customers of the Parts and Service departments that are likely to leave so that an adequate retention campaign may be planned. The problem area is therefore the customer relationship management which involve all business processes and departments. As such, the secondary goal is determining the set of variables from the Customer Data (the information kept regarding the customer from a business perspective) that are critical for retention. To achieve this, the framework presented in the literature review (**Figure 2.8**) is applied: customer loyalty is affected by supporting and repressing factors and the challenge of the data miner is to define them in rigorous terms.

Loyalty-supporting factors causing dedication are the first to be analyzed. Dedication means the will of a customer to be involved in a long term relationship with the Client, therefore an important aspect is the duration of the relationship. This can be derived from a first date there has been an interaction between the customer and the Client and may be the date when the customer was registered in the system, the date the first order was placed, the date the first delivery was made or the invoice was issued. The decision taken together with the Client is to use the date of the first order since it is a valid document and it shows a commitment from the customers before even paying any amount to the Client. Other important information that shows satisfaction and trust between the customers and the Client is the number of orders placed since the start of relationship. Based on this concept several important derived information can be determined.

The number of orders and the total amount of the orders require a calculation by months or quarterly in order to show trends, moreover the orders need to be divided by type: Parts or Service. A Parts order may have many items of type spare parts while Service order may have many items of type spare parts and service. The number of monthly items (in general called transactions of the order) is also a good understanding of the customers' behaviour: it is useful to find out if the customer buy few items of high value or many items of low value.

Each item (spare parts or service) belongs to a material category. The material category is a way of grouping items together and defining for this group a set of properties like the formulae to calculate the *EUP* (end user price). Usually the formula for the *EUP* has the following components:

$$EUP = BASE * (1+Packaging+Transport)$$

While the *BASE* price is the price of the item on the supplier catalogue, therefore there is a specific price for each item, the *Packaging* and *Transport* constants are properties of the material category.

Material Category	Transport	Packaging	EUP Formula
[P] Filter, belt, batteries, etc...	0.05	0.025	$BASE * (1+Packaging+Transport)$
[Y] Workshop tools	0.05	0.035	$BASE * (1+Packaging+Transport)$
[CS] Service	0	0	$BASE * (1+Packaging+Transport)$

**Table 4.1:** example of material categories

As per **Table 4.1**, say an item costs 100 (*BASE* price) and it belongs to the category [P], the *EUP* price for the customer is:

$$EUP = 100 * (1+0.05+0.025) = 107.5$$

Services (belonging to categories like category CS) obviously do not have any transport or packaging factor that alters the price. On the other hand, the transport and packaging constants for spare parts depends mainly on the dimension of the items.

Because the material category affects the price of the transaction it is also wise to distinguish the number of monthly transactions by material category.

Finally, the transactions are also characterized by the delivery priority which relates to the choice of the customers in having the items delivered earlier (priority = URGENT) than the standard estimated delivery (priority = STOCK).

Nordman (2004) reports different type of Loyalty-supporting factors of type constraints of which the relevant in this case study are: economic bonds, lack of alternatives, geographic bonds.

Economic bonds concern with the benefits that a customer lose when switching to a new provider.

The Client has three customer categories (summarized in **Table 4.2**) to manage different pricing (hence benefits) for each customer depending on the type of order placed and the supplier of the items chosen.

Customer Categories	Supplier	Sale Organization
A	Supplier I	Parts
B	Supplier I	Service
C	Supplier V	Parts and Service

**Table 4.2:** customer categories applied to placed orders

In practice, if a customer places a Parts order with parts from the Supplier I, the pricing applied to the items is calculated based on the customer categories A. If instead, for example, a customer place a Service Order, whereas the parts ordered are from Supplier V, the category applied is C.



**Table 4.3** shows for each customer category the list of categories (not exhaustive) that can be associated to a customer whilst **Table 4.4** reports an example of association between customers and categories.

Customer Categories A	Customer Categories B	Customer Categories C
VIP (24%)	PLATINUM14 (14%)	VIP (24%)
GOLD card (10%)	PLATINUM15 (15%)	VIP10 (10%)
WELCOME (3%)	DIAMOND(10%)	DEALER(3-32% depending on material category)
NEW SUBDEALER (7%- 28% depending on material category)	GOLD(7%)	DEALER-10(10%)
SUBDEALER(32%)	SILVER(3%)	DEALER-12(12%)
	WELCOME(3%)	

**Table 4.3:** list of categories of each customer category and their percentage discount

Customer Identifier	Customer Categories A	Customer Categories B	Customer Categories C
C32826	NON-CATEGORIZED	DIAMOND	VIP
C5457	SUBDEALER	NON-CATEGORIZED	VIP
C727	NON-CATEGORIZED	GOLD	VIP

**Table 4.4:** example showing categories assigned to specific customer

Continuing with the example above, if a customer place a Parts order with parts from the Supplier I, the pricing applied to the items is calculated based on the customer categories A. In case the customer is C5457, the category A used is SUBDEALER which means 32% discount on the item price (to be specific on the *EUP* price which depends on material category introduced above).

Having introduced how customers are categorized, it is clear that for economic bonds, the supporting factors are the customer categories A, B and C associated with the customer.

Another economic bond factor worth mentioning is the credit facility: some customers are entitled to use credit while others are not (must pay by cash).

To establish if the customer is staying due to lack of alternatives, an analysis of the competitors is required. Competitors are dealerships providing Service and Parts from the same suppliers (identified as Supplier I and Supplier V). In **Table 4.5** the competitors and the Client geographic areas are presented (in Data Understanding stage is shown how areas are associated).

Company Name	Area
Client	Area177,Area178,Area179
Competitor1	Area681,Area682,Area683,Area684
Competitor2	Area236,Area237,Area238,Area239,Area240, Area241,Area242,Area243,Area244,Area245, Area246,Area247
Competitor3	Area236,Area237,Area238,Area239,Area240, Area241,Area242,Area243,Area244,Area245, Area246,Area247
Competitor4	Area718,Area719,Area720
Competitor5	Area1405,Area1406,Area1407,Area1408, Area1409,Area1410,Area1411,Area1412
Competitor6	Area1405,Area1406,Area1407,Area1408, Area1409,Area1410,Area1411,Area1412
Competitor7	Area715

**Table 4.5:** *area where Client and competitors run their business*

By comparing the customers' area with the competitors' area a variable can be defined to determine if the customer has or hasn't lack of alternative.

Finally, for geographical bonds, meaning that the location ties the customer to the provider, the area of the customers must be compared to the location of the Client, similarly to the lack of alternative whereas the area of customers is compared to the area of competitors.

Repressing factors, as already described in the literature review, are not directly the opposite of supporting factors. One of the main aspects that negatively influence the loyalty status is related to pricing which means, in the case under study, anything that affects the *BASE* price. The investigation carried out with the help of the Client on the supplier catalogue gave the following results: the *BASE* price for almost all items change every 6 months, while every week they may be price adjustment for a small subset of items. For this reason, a critical variable to take into account is which item is bought by the single customer the most (in the period investigated) and if that item has been increased or decreased in price since the period before. It may be that the reason of the churning is due to the fact that the orders amount became too expensive as opposed to the previous periods. But it is also fair to say that this may not be the cause if there is an alternative item to the one normally bought which is cheaper. Usually an item is sold from the OEM (Supplier I) but alternatively it may be substituted by the same item sold by Supplier V. For that reason, a variable considering this information on alternates is necessary.

Another repressing factor is the lost of benefits: for example a customer having a VIP customer category A is changed back to GOLD card. They may be valid reason of why the Client made this choice but it is useful to understand if this may impact the churning more than other variables. The Client

keep a history of the customer categories hence it is easy to highlight if there has been a change from one quarter to another in terms of discounts.

To conclude on repressing factors, the quality of service need to be somehow quantified. A direct measure is the number of days the service has taken to repair a vehicle and it is the difference between the date the order is place and the completion date when the truck is given back to the customer.

The variables identified so far were all based on the framework proposed regarding supporting and repressing factors of the loyalty status. In classification algorithms socio-demographic information are often included (Giudici, 2003). The information the Client thinks to be representative are: Area, Country and Validity. The Validity is an indicator on the financials details of the customers. The indicator is true not only if all financial data is filled but also if the fiscal code is validated against the European Commission Taxation and Customs Union web site.

The Business Understanding stage, if on one side focus on the business background and the business primary and secondary goals of the mining project (as described above), on the other side produce an additional essential output which is the project plan. This has been part of the project definition and added in the Appendix A.

## 4.2 Data Understanding

The Client runs the business processes using an ERP system, hence the data to be analyzed is stored in a single centralized database. The relational database management system (RDBMS) adopted for the ERP system is MySQL and the data, for the purpose of this stage and the next Data Preparation stage, is accessed by using the MySQL client command line and the SQL query language. Whereas needed, the data from MySQL database is outputted in a text file and loaded as a matrix variable in either Matlab or Excel in order to perform graphical visualization and analysis.

The main output of this stage is the selection of tables (datasets), rows (instances) and columns (attributes), their descriptions and eventually graphical data explorations. This stage, together with the Modelling stage, represents the findings of the project; in fact not only it is important the final data mining model but also the in depth analysis of the data that produced such as model.

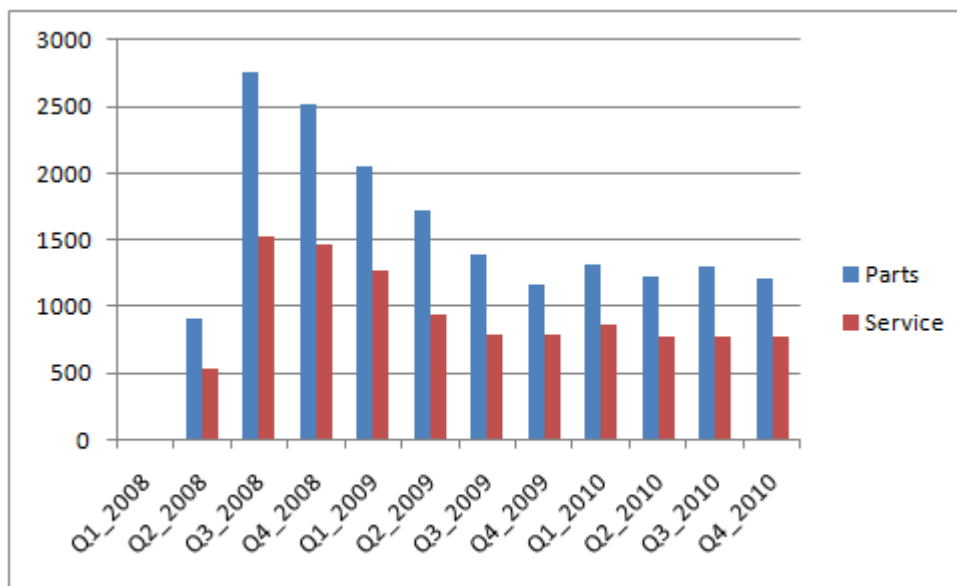
The starting point is the list of variables identified with the Client in the Business Understanding stage, see table in Appendix B. There are in total 65 explanatory variables and 1 target class: those are analysed and discussed one by one in the order reported in Appendix B.

The first group of explanatory variables concerns the Parts and Service orders information: first order of the customer, counts and amounts of Parts and Service orders per customer by quarter. This information can be retrieved from the ORDER table (see Appendix C for the tables and fields selected from the Client database), in particular looking at the order date field, the order amount and currency and the order type.

**Figure 4.2** shows the numbers of orders in the ORDER table grouped by the categorical attribute order type (Parts or Service) and the attribute order date which has been subdivided in 4 quarters (Q1, Q2, Q3, Q4) applying a data discretization method (see **Table 4.6**).

Quarter	Order date from	Order date to (not inclusive)
Q1	(Year)-01-01	(Year)-04-01
Q2	(Year)-04-01	(Year)-07-01
Q3	(Year)-07-01	(Year)-10-01
Q4	(Year)-10-01	(Year+1)-01-01

**Table 4.6:** data discretization of the attribute order date

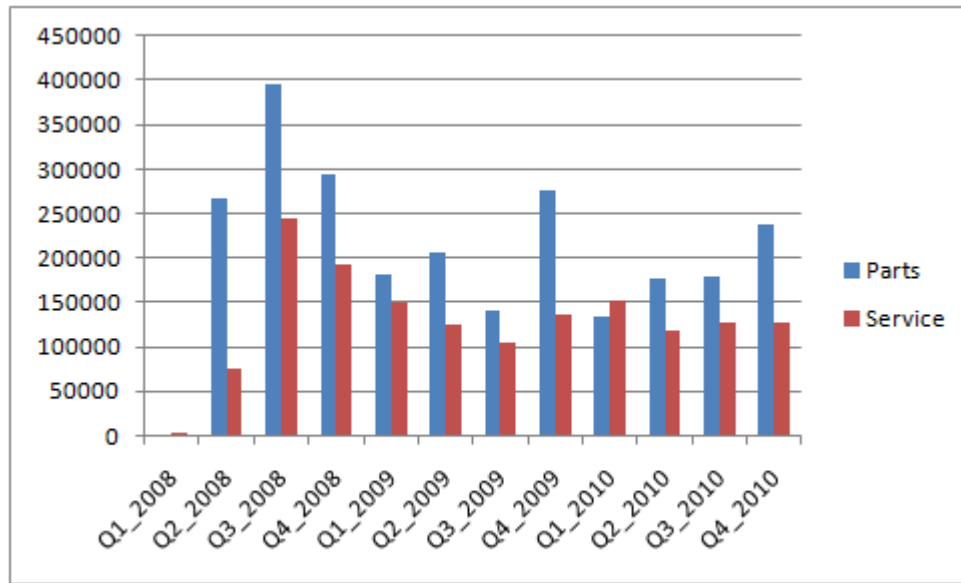


**Figure 4.2:** number of orders by quarter and order type

Quarterly, on average, the customers place around 1500 orders of type Parts and 900 orders of type Service. The fact that in Q1 2008 there are 0 orders and in Q2 a very low number of orders is due to an internally agreed procedure that archives all the data from the current ERP system database to another database (the data warehouse). The obvious advantage of such operation is to have a faster operational system while keeping (legally mandatory for financial audits) the older data to a different database. The missing information from 2008 and previous years is not important for the purpose of calculating the orders amount and counts explanatory variables since the focus of this analysis is on Q1, Q2, Q3, Q4 2010. As explained in more details in the following Data Preparation stage, the data period analyzed is 2010 which is used to predict 2011 and then use the data from 2011 to validate the results.

However, the data from the data warehouse is useful to calculate the first explanatory variable which rely on the first order date (Days from First Order Date). Some of the customers have been loyal since years prior to 2008.

In **Figure 4.3** a similar vertical bar chart, this time showing the orders amounts by quarter.



**Figure 4.3:** orders amounts by quarter and order type (values in euro)

The average order revenues by quarter are 220.000 euro for Parts order and 140.000 euro for Service. Those two charts give an idea of the number of records (instances) in the table ORDER as well as an idea of the financials of the Client in terms of revenues trends over the last 3 years.

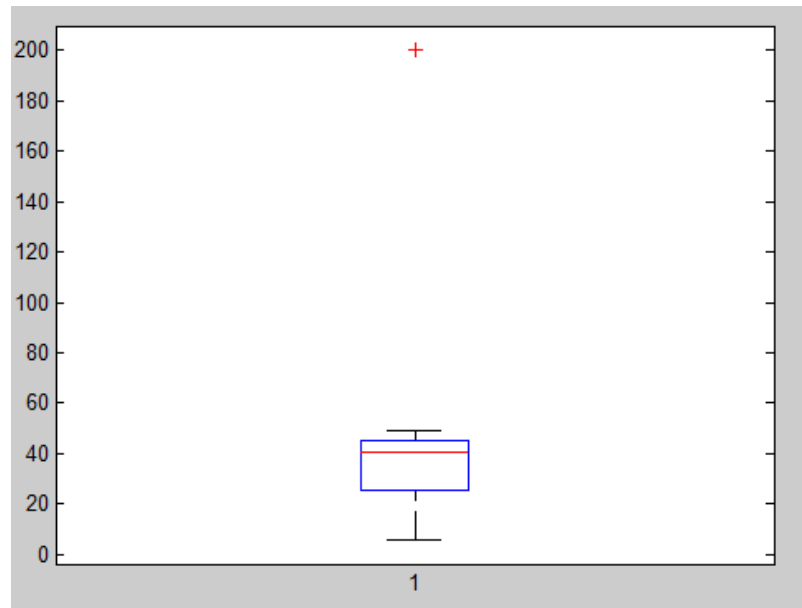
It is critical at this stage to verify the quality of the data examined so that the data mining algorithms are able to produce the correct answers. A common exercise is to identify anomalies or so called outliers, defined as observation that are significantly distant from the rest of the data (Barnett and Lewis, 1994). In other words, it is crucial to establish if the figures calculated in **Figure 4.3** are affected by customers placing significantly big orders, which, as such, are anomalies in the sense that are special customers and should not be considered as part of this analysis. The aim of the project is to find rules that apply to most of the customers and those rules should not be affected by exceptional customers for which the Client has already dedicated channels.

A way of identifying outliers is using a box plot, also referred as box-and-whisker plot, which can be considered a visual representation of the quartiles and the *IQR*. To keep it simple, the quartiles are the three values (lower quartile *Q1*, median *Q2* and upper quartile *Q3*) that divide the dataset *T*, which has been ordered by ascending order, into 4 equal groups (25% of the population in each group). The *IQR* (interquartile range) is the different between the lower and upper quartile ( $IQR = Q3 - Q1$ ).

The box plot (see **Figure 4.4**) shows the median, the quartiles and the lower and upper edge calculated as:

$$\text{Lower Edge} = Q1 - 1.5IQR$$

$$\text{Upper Edge} = Q3 + 1.5IQR$$

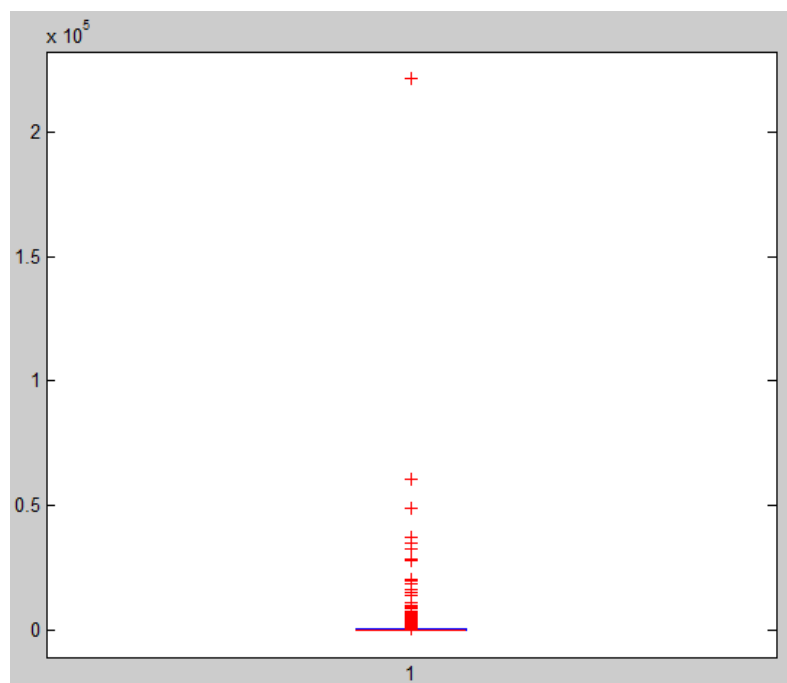


**Figure 4.4:** example of box plot using Matlab, outliers marked with a plus symbol

The values that are outside the lower and upper edge can be considered outliers.

Applying the same concept to the dataset under analysis, where the single observation is the total order amount generated by each customer over the period considered (2010), it gives the results in

**Figure 4.5.**

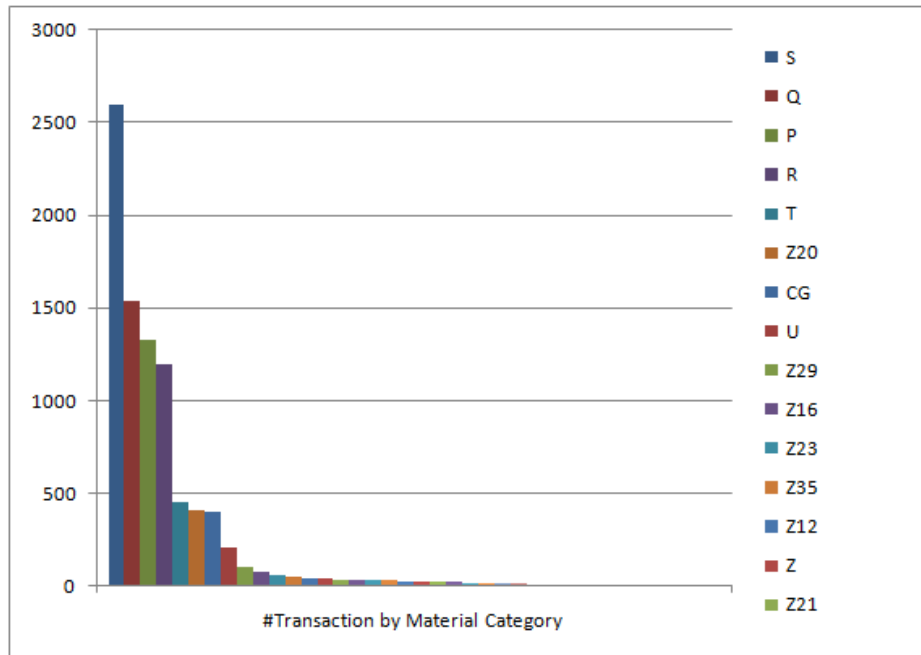


**Figure 4.5:** box plot applied to the customers total order amount in 2010

It is clear from the box plot that there is an exceptional customer (CUSTOMER655) ordering over 200.000 euro and few customers (to be exact 8) ordering more than 20.000 euro (note that the number on the  $y$  axis are multiplied by  $10^5$ ). Those are the customers excluded from the analysis. CUSTOMER655 have such an anomalous behavior because it is a company of the same group of which the Client is part of: the Client buy at its market (country) prices to resell parts to the peer company of another market.

The second group of explanatory variables (again, see order in Appendix B) involves the analysis of the number of transactions per customer by quarter and consequently grouped by material categories and delivery priority. The relevant table is the TRANSACTION table and the fields considered are Material Category and Delivery Priority stored for each transaction (each item on the order).

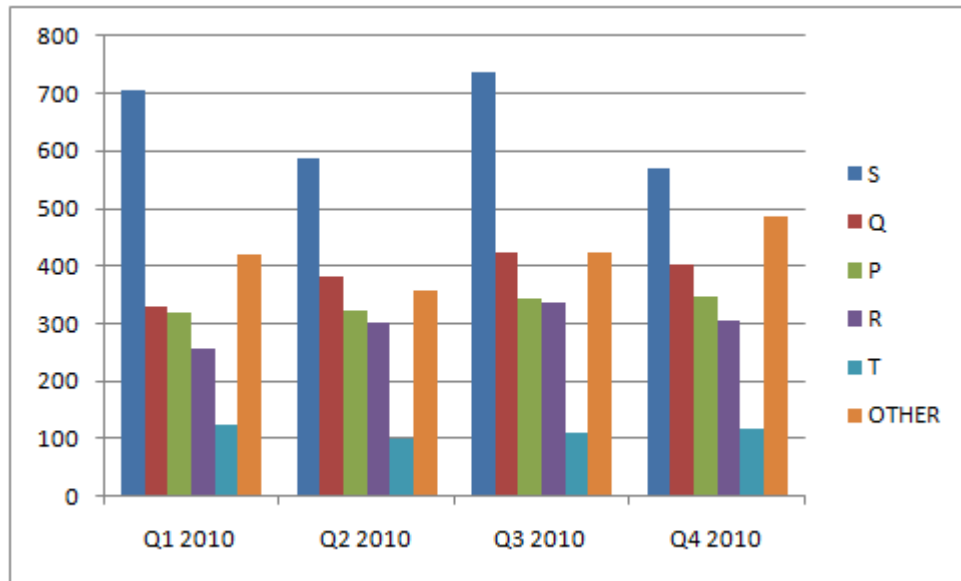
First of all, the top 5 material categories purchased by the customers in 2010 need to be delineated. This is done by drawing an histogram of the number of occurrences (transactions) by all material categories (**Figure 4.6**).



**Figure 4.6:** top 5 material categories

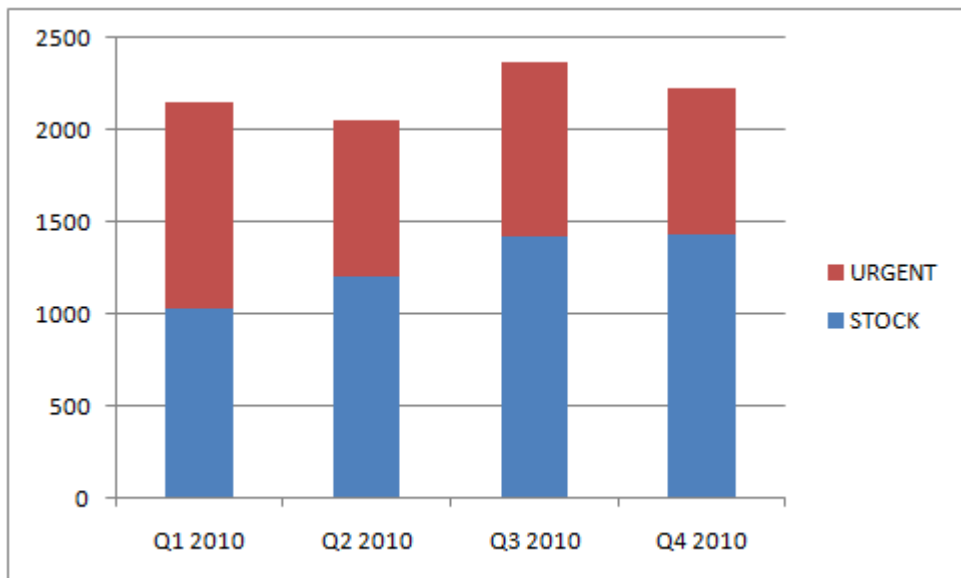
The top 5 material categories are: S, Q, P, R, T which correspond to MatCat1, MatCat2, MatCat3, MatCat4, MatCat5 labels given to the explanatory variables in Appendix B. All others material categories are grouped together in a single category called OTHERS (MatCatOthers).

Once the material category is defined, the number of transactions by quarter and top 5 material categories are shown in **Figure 4.7**.



**Figure 4.7:** transactions count in 2010 by material category

Same number of transactions, grouped by quarter and delivery priority (STOCK or URGENT) are shown in **Figure 4.8**.

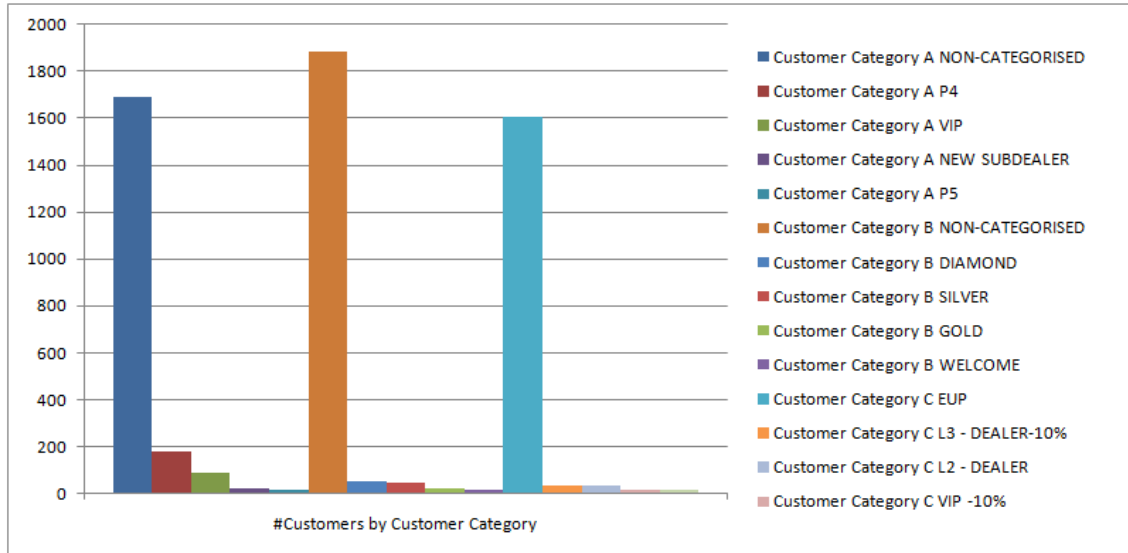


**Figure 4.8:** transactions count in 2010 by delivery priority

It is worth noting that the number of urgent and stock transaction per quarter are comparable numbers and it can potentially be another rule that separates churning and non-churning behaviours. A final consideration on transaction counts is that the number of transactions per quarter is also comparable to the number of orders per quarter, which implies that each order only contains a low number of items.



The explanatory variables Customer Category A, B and C, Credit Indicator are a simple select from the CUSTOMER RULE table. **Figure 4.9** shows how the customer categories categorical attributes are distributed.



**Figure 4.9:** top 5 customer categories for A, B, C customer category

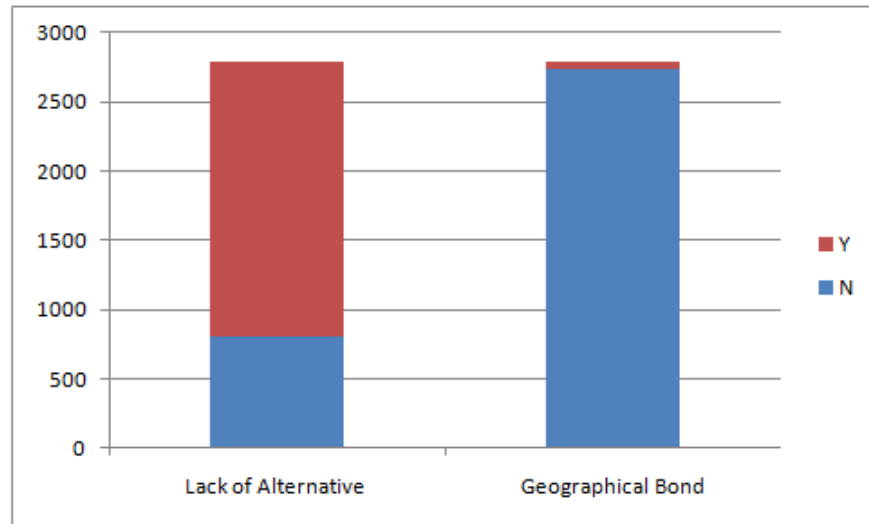
The result of this investigation say that only a small number of customers have properly assigned categories, the majority have the generic category NON-CATEGORISED which means no discounts are applied to the *EUP* (end user price). This could be one of the cause that affect the churners: if that's the case, the obvious action is for the Client to start better differentiating the way the customers are treated as individuals. The Modelling stage answers this question.

**Figure 4.9** also highlight another important information which is the total number of customers under investigation that is just below 3000 (summing all the bars for a single customer category, in figure only shown the top 5). In other words, the final dataset *T* for the mining algorithms have 3000 rows (instances or customers) and 65 columns (explanatory variables for each customer defined in Appendix B).

A final note on the CUSTOMER RULE table is that of those 3000 customers, 20% has got credit facility (explanatory variable Credit Indicator).

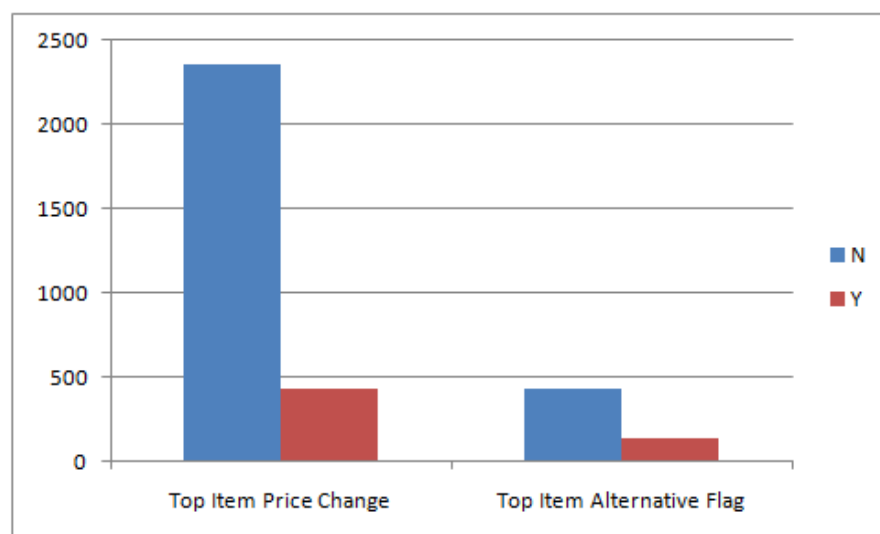
The last two variables of type Supporting-factor to be discussed are the lack of alternative indicator and the geographical bond indicator, both based on the Area field from the CUSTOMER table. Area is a categorical attributes assuming 1600 distinct values, many of them duplicates referring to the same area due to inconsistencies: spelling mistakes, upper and lower cases, abbreviations, area concatenated with the street name. This is the reason why, as reported in **Table 4.5**, the Client area as well as the competitors areas correspond to a list of area ids and not just a single id. If the customer area id is one of the list corresponding to the Client area, the indicator geographical bond is true while if the customer id is not one of the competitors area list, the indicator lack of alternative is true. **Figure 4.10** illustrates that most of the customers hasn't got a geographical bond meaning that they perform

their business activities in areas different than the Client. On the other hand there are a substantial percentage of customers that have alternatives and this may affect the churning decision.



**Figure 4.10:** customers' area compared to Client and competitors' area

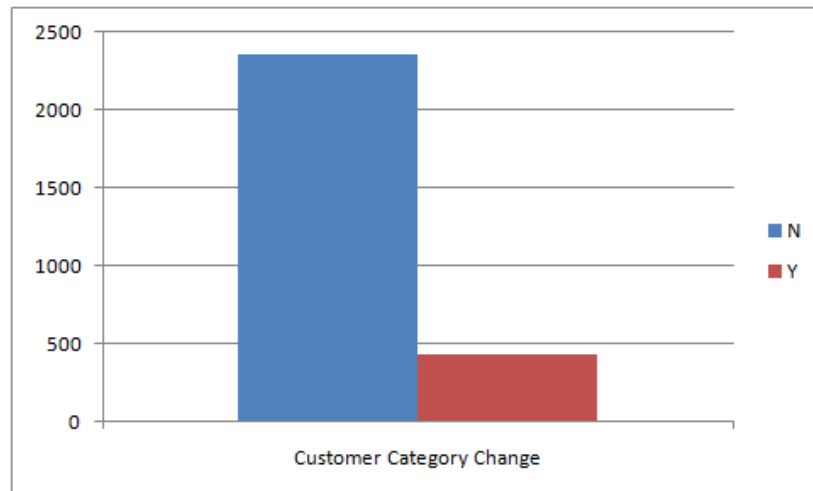
The variables of type Repressing-factor also are relevant in understanding customer retention. The first two variables of this type are based on the top item purchased by the customer in the period investigated. Once the item has been found, the 2010 price of the item is compared to 2009 price of the same item (ITEM table): if 2010 price is greater than 2009 price, the explanatory variable Top Purchased Item price change is set to true. Moreover, the related variable Top Purchased Item alternative flag is set to true if for this item there is an alternative (an equivalent part produced by another Supplier, ITEM ALTERNATIVE table). The results are shown in **Figure 4.11**.



**Figure 4.11:** customer top bought item price change and alternates

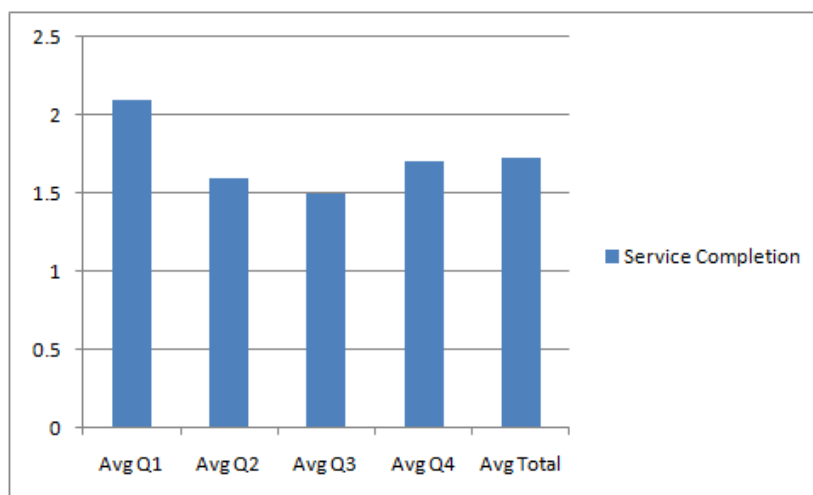
Less than 500 top items had a price change and of those about 25% have an alternative item.

**Figure 4.12** gives an idea of the number of customers had changes in terms of discounts (positive or negative) from 2009 to 2010. If there was a change, the explanatory variable Customer category change indicator 2010 is set to true.



**Figure 4.12:** changes in customer categories from 2009 to 2010

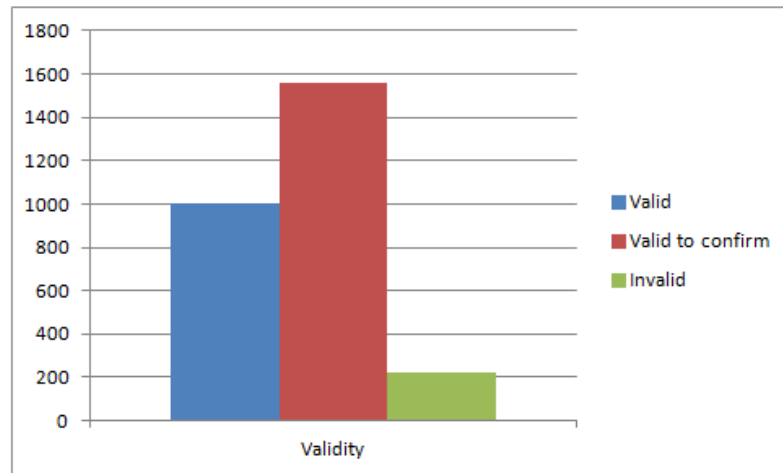
Finally, a bar chart (**Figure 4.13**) highlighting the average service completion by quarter (explanatory variables Average Service Completion days Q1, Q2, Q3, Q4) querying the VEHICLE ORDER table.



**Figure 4.13:** service completion average by quarter as measure of quality

The average total is 1.7 days (rounding up, 2 days), which means that the customer have the expectation that the Client repairs take no more than 2 days.

To conclude, socio-demographic data Area, Country and Validity are straight select from the CUSTOMER table. The majority of the customers (94%) are from the same country of the Client. Area has been already discussed (**Figure 4.10**) while the analysis on Validity is presented in **Figure 4.14**.



**Figure 4.14:** *customer validity indicator*

Valid to confirm are the customers having all the required financial data inserted in the system. Valid are the customers having all required financial data but also manually confirmed by the personnel. Invalid are customers without some of the important financial information like the fiscal code or fiscal address.

### 4.3 Data Preparation

In Data Understanding stage the relevant tables and columns have been selected and graphically explored. Fayyad et al. (2002) describe the importance of data visualization in a data mining project and provide a visualization-driven approach for strategic knowledge discovery.

The aim of the Data Preparation stage is more practical: construct one final derived table (final dataset  $T$ , fields listed in Appendix B, SQL queries reported in Appendix F) consisting in all the variables individually analyzed in the Data Understanding stage. Because the data mining algorithms presented in the literature review are supervised learning process, not only it is necessary to provide a list of explanatory variables but also the value of the target class (Churner indicator) for the provided instances.

Recognizing churn is not a straightforward task in the sense that it differs from business to business: it is certainly easier when there is a monthly billing relationship since if the payments are stopped the customer can be considered a churner (Berry and Linoff, 2004). The Client hasn't got for all customers this type of relationship: only 20% is allowed to have credit, the rest pay by cash. For this reason a different definition of churner needs to be contemplated.

For the time frame 2010, the total order amount purchased by each customer is calculated as:

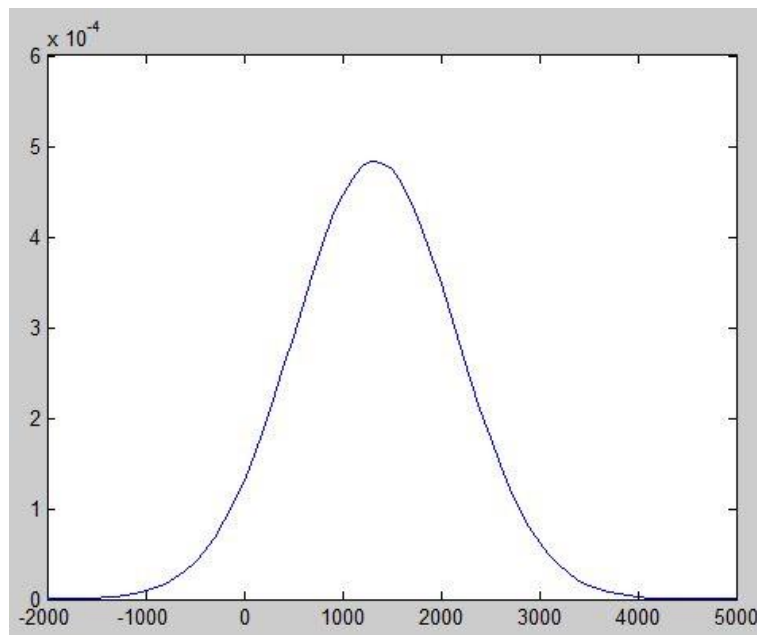
$$\begin{aligned}
 \text{Customer Total Order Amount 2010} = & \\
 & \text{Parts Order amount Q1 2010} + \\
 & \text{Parts Order amount Q2 2010} + \\
 & \text{Parts Order amount Q3 2010} + \\
 & \text{Parts Order amount Q4 2010} + \\
 & \text{Service Order amount Q1 2010} +
 \end{aligned}$$

$$\begin{aligned}
 & \text{Service Order amount Q2 2010} + \\
 & \text{Service Order amount Q3 2010} + \\
 & \text{Service Order amount Q4 2010}
 \end{aligned}$$

First of all, given the yearly totals, it has been agreed to filter the dataset  $T$  for the only instances where the sum is over 500 euro. In other words, if a customer spent at least 500 euro in a year, it is worth retaining from a business perspective.

Data filtering is another important component of this stage, the aim is to have rules from the mining algorithms that apply to the right subset of customers and not all of them (in fact, excluding outliers and not valuable customers). The dataset  $T$  originally containing about 3000 instances (to be accurate 2785 rows), once filtered, only hold 286 samples. **Figure 4.15** shows that on average (excluding outliers) the total order amount is around 1300 euro (the mean  $\mu$ ), with a standard deviation  $\sigma$  about 800 euro.

The interval  $\mu \pm \sigma$  contains approximately 68% of the possible values (Vercellis, 2009; Grinstead and Snell, 1997). Strictly speaking, 68% of the customers purchased between 500 euro (1300-800) and 2100 euro (1300+800) worth of goods in 2010.



**Figure 4.15:** normal distribution of the customer total order amount 2010

Given the customer amount ordered in Q1 2011 (future period compared to the one analyzed):

$$\begin{aligned}
 & \text{Customer Order Amount Q1 2011} = \\
 & \text{Parts Order amount Q1 2011} + \\
 & \text{Service Order amount Q1 2011}
 \end{aligned}$$

the target class (Churner indicator) reflects the following condition (applied to the filtered dataset  $T$ , so customers purchased at least 500 euro in 2010):

*IF (Customer Order Amount Q1 2011=0 and Customer Order Amount Q1 2010>0)*

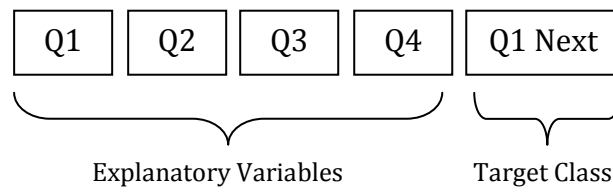
*Churner indicator = TRUE*

*ELSE*

*Churner indicator = FALSE*

For example, a customer purchasing 1000 euro in 2010 (more than 500 euro which means valuable customer) having purchased a certain portion of the 1000 euro in Q1 2010 but nothing in Q1 2011, can be considered a churner. Or, to put it in another way, if the customer used to buy every quarter on average 250 euro ( $1000/4$ ) and didn't buy anything in the first quarter of the future period (Q1 2011) is likely to have left the Client.

A final consideration of the time frames used is required to understand prediction algorithms. As represented in **Figure 4.16**, the explanatory variables defined in the Data Understanding stage are from a period (2010) previous to the target variable (2011): the aim is to find patterns from one period to explain outcomes of a later period (Berry and Linoff, 2004). In practice, the algorithms use the period 2010 as training data and 2011 as target data.



**Figure 4.16:** predictive data mining time frames

To summarize, after performing the validation and filtering tasks mentioned above, the final dataset  $T$  used as input to the next Modelling stage contains 66 columns (65 explanatory and 1 target) and 286 rows.

## 4.4 Modelling

Hadden (2008) and Mutanen (2006) conducted a survey on the number of papers and related algorithms used for customer prediction concluding that the most used are: Decision Tree, Logistic Regression and Neural Networks. Those algorithms have been formally introduced in the literature review and are applied to the dataset  $T$  prepared in the previous stages of the project.

The activities of this stage consist in: build the 3 models based on the 3 type of algorithms separately, for each of them describe the results and assess the accuracy, finally compare their accuracy and determine which the best predictor is.

The very first step, common to all models, is to load the dataset  $T$  into Weka machine learning software. As shown in **Figure 4.17** this is obtained by connecting the tool directly to the MySQL database and selecting the information from the denormalized final table that has been created in the Data Preparation stage (see Appendix F).

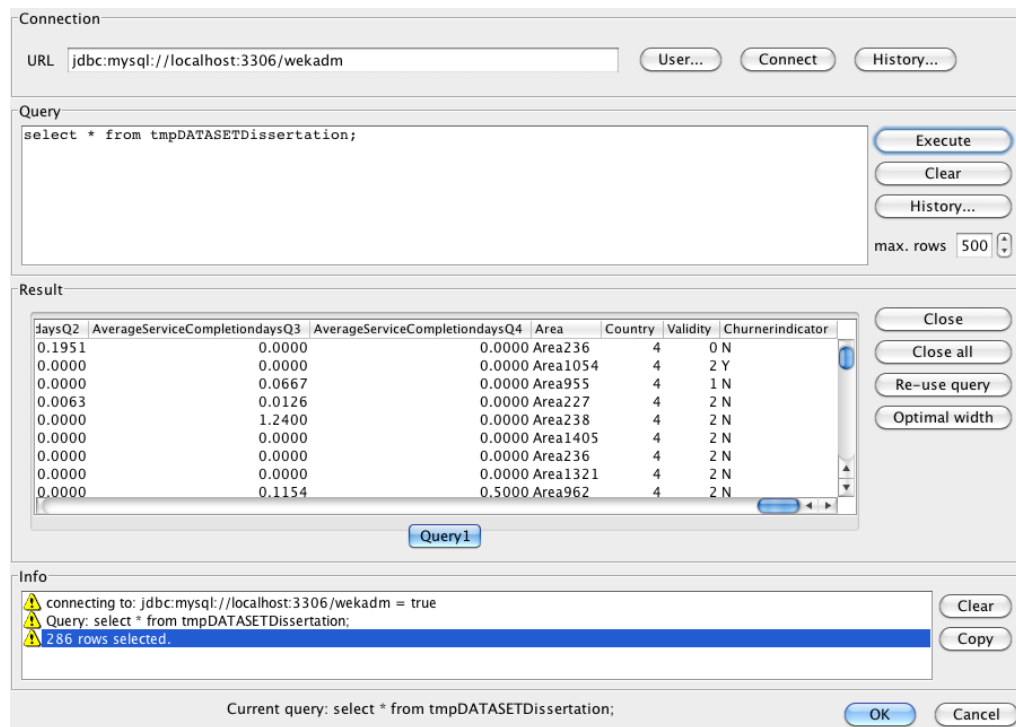


Figure 4.17: loading data into Weka software

Weka allows different type of data source: natively support the ARFF format (attribute-relation file format) but there are also available converter for spreadsheets (cvs files) and XRFF (xml attribute relation file) and direct SQL connection to the database (method used in this project).

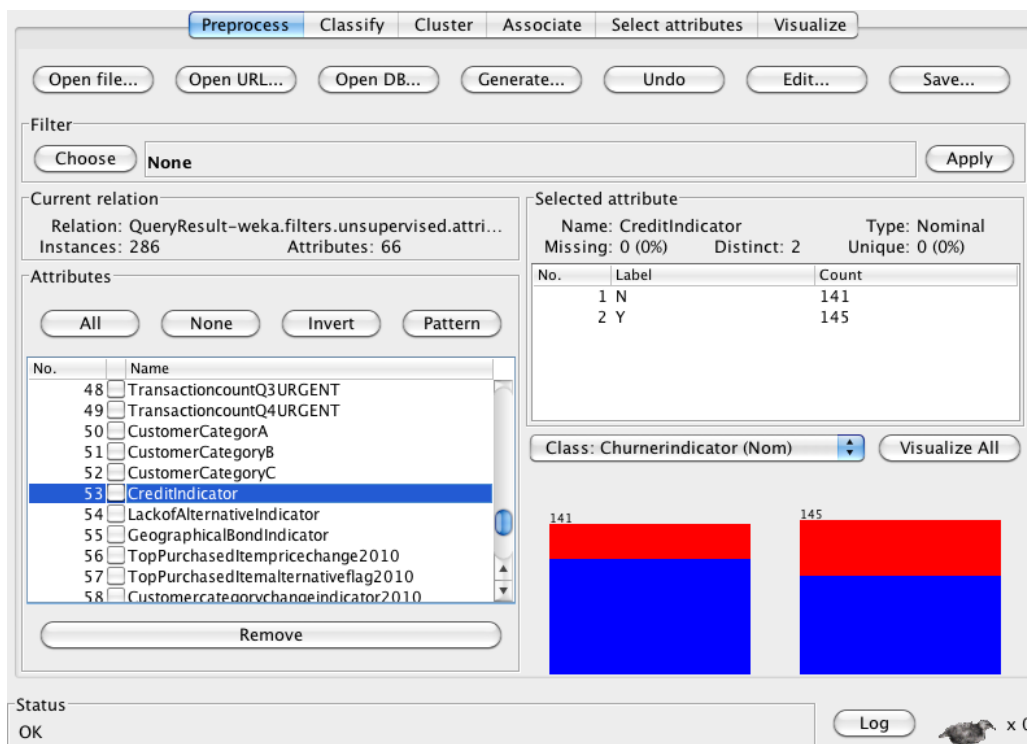
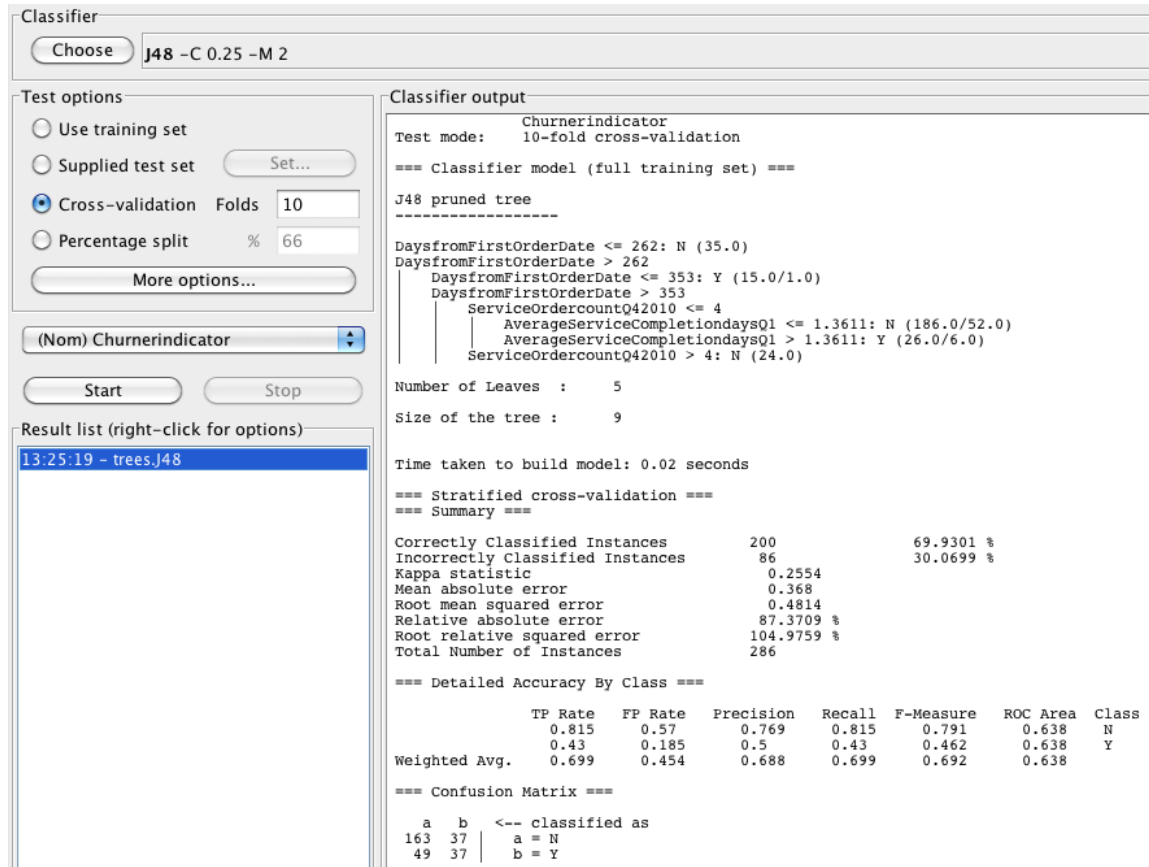


Figure 4.18: preprocessing interface showing the dataset  $T$  summary

Once the dataset  $T$  is loaded, the preprocessing interface shows key information on the data: there are 286 instances and 66 attributes, as expected (**Figure 4.18**). Moreover, it gives the miner the ability to perform bivariate analysis. So far, in particular in the Data Understanding stage, it has been conducted only analysis of type univariate, meaning that only one attribute at a time has been explored. Bivariate analysis considers two attributes (in this case an explanatory attribute and the target class Churner indicator) and the relation between the two (Mardia et al., 1979). For instance, in **Figure 4.18**, the categorical attribute Credit indicator has been selected among the list of available explanatory variables. The bar chart tells that there are 141 instances having the Credit indicator to N (false) and 145 to Y (true). Considering that the red part of a single bar means churner, one may say, surprisingly, that there are more churner in the category of customers having credit facilities than to the category of customers that haven't. The findings of the modelling give a better understanding on which of the explanatory variables it is useful to conduct this type of bivariate analysis.

The first classification task is performed adopting the Decision Tree algorithm. Weka provides different implementation of decision trees: J48, SimpleCart (minimum cost-complexity pruning), REPTree (reduced-error pruning). J48 is the implementation of the C4.5 algorithm developed by Quinlan (1993) which is the extension of the algorithm presented in the literature review adapted to real world problems. C4.5 is the most widely used algorithm in machine learning projects (Witten et al., 2011), hence the one chosen.



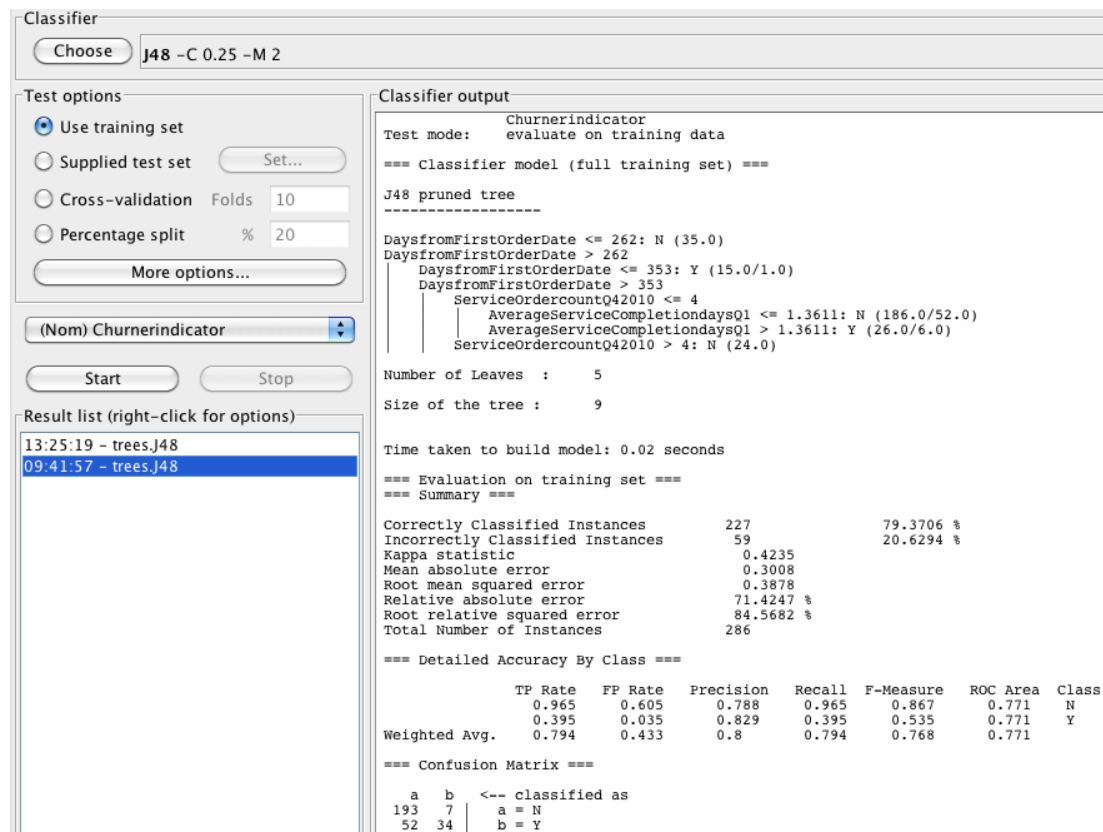
**Figure 4.19:** C4.5 (J48) classifier results on dataset  $T$



**Figure 4.19** presents the results of the first model which uses the decision tree method. Before discussing the output of the decision tree, it is necessary a digression on the test option used: the 10-fold cross validation (which is test option selected also for the other 2 models).

In the literature review it has been mentioned that to evaluate the accuracy of the model, the dataset  $D$  is split in two disjoint subset: the dataset  $T$  to train the model and the dataset  $V$  to evaluate it. The problem with this general and theoretical approach is that the accuracy of the model depends entirely on the instances randomly selected for the dataset  $T$ , which may be not representative for a class: if all the instances of this class are by chance only in the dataset  $V$ , this class is not included in the model at all (Witten et al., 2011). This behaviour is more prominent in small dataset like the one created for this project (286 instances) and it is wise to use the whole set to train the model and not part of it. For those reasons the 10-fold cross validation option has been chosen. The idea is that the dataset  $T$  (286 instances) is divided in 10 disjoint subsets and the procedure of training the model and evaluating the accuracy is repeated 10 times: each time 1 subset is selected for evaluation and the union of the other 9 subsets for training. At the end of the 10 iterations, the final accuracy of the model is the average of the accuracy of each single procedure. This ensures that each instance appears once in the valuation set and the same number of times in the training set, making the model more robust, in other words the results do not vary significantly as the instances varies in the two sets (Vercellis, 2009).

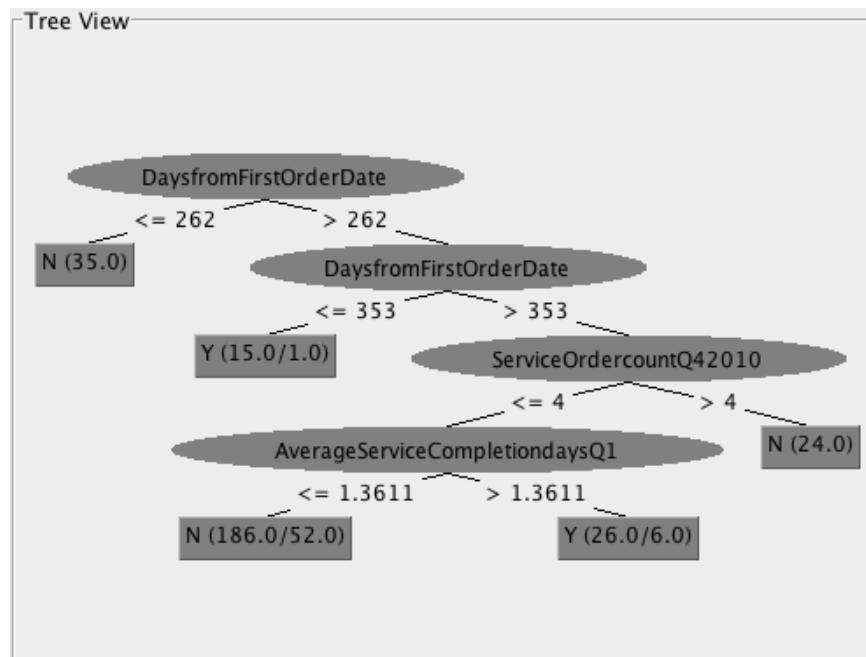
Having run the 10-fold cross validation technique, the decision tree model generated using the dataset  $T$  classifies correctly 69.93% of the instances and incorrectly classifies 30.07% of the instances: in practice, of the 286 instances, 200 correctly classified and 86 incorrectly classified (**Figure 4.19**).



**Figure 4.20:** optimistic decision tree model

**Figure 4.20** confirms that running the model using as test option the entire training set (without folds validation) gives a better accuracy results (227 correctly classified and 59 incorrectly classified) but suffers of being less generic and too specific to the dataset, strictly speaking not a good predictor and not a robust model.

The decision tree resulting from the model is presented in **Figure 4.21** (it is also shown in its text version in **Figure 4.19**).



**Figure 4.21:** *Weka Classifier Tree Visualizer interface*

Accordingly to the model, the first and second rule are based on the Days from First Order Date explanatory variable. If the customer has got a relationship with the Client that is less than 262 days, it is classified as non churner; if greater than 262 but less than 353 days (less than a year) it can be a potential churner; if greater than 353 then it depends on the number of service orders placed in the last quarter under analysis (Service Order count Q4 2010). More than 4 service orders in Q4 the customer is not a churner, less or equal 4 is a churner or not depending on the quality of service the customer had in Q1 (Average Service Completion days Q1). On the leaf of the tree, 26.0/6.0 means that of 26 instances of the class Y (churner) 6 were incorrectly classified as such.

In the next Evaluation stage the results from all models are discussed in the business context and compared to the objectives of the data mining project, for this stage only the findings are reported.

The results from the second model are reported in **Figure 4.22**.

```

Class 0 :
0.65 +
[PartsOrdercountQ32010] * 0.05 +
[PartsOrdercountQ42010] * 0.04 +
[ServiceOrdercountQ12010] * -0.06 +
[ServiceOrdercountQ32010] * 0.11 +
[ServiceOrdercountQ42010] * 0.08 +
[PartsOrderamountQ32010] * 0 +
[ServiceOrderamountQ12010] * 0 +
[ServiceOrderamountQ42010] * 0 +
[TransactioncountQ4MatCat1] * 0.08 +
[TransactioncountQ4MatCat2] * 0.08 +
[TransactioncountQ1MatCat5] * -0.35 +
[TransactioncountQ4MatCat5] * 0.31 +
[TransactioncountQ1STOCK] * -0.03 +
[CustomerCategoryB=NON-CATEGORISED] * -0.69 +
[CustomerCategoryC=L2 - DEALER] * 0.71 +
[CreditIndicator] * -0.27 +
[GeographicalBondIndicator] * -0.66 +
[Area=Area333] * -1.51 +
[Area=Area368] * -1.51 +
[Area=Area550] * -1.15 +
[Area=Area1502] * -1.51 +
[Area=Area1276] * -1.51 +
[Area=Area728] * 1.23 +
[Area=Area114] * -1.51 +
[Area=Area28] * -1.51 +
[Area=Area1440] * -1.51 +
[Area=Area685] * -1.51 +
[Area=Area1546] * -2.61 +
[Area=Area964] * -1.51 +
[Country] * 0.12

Class 1 :
-0.65 +
[PartsOrdercountQ32010] * -0.05 +
[PartsOrdercountQ42010] * -0.04 +
[ServiceOrdercountQ12010] * 0.06 +
[ServiceOrdercountQ32010] * -0.11 +
[ServiceOrdercountQ42010] * -0.08 +
[PartsOrderamountQ32010] * 0 +
[ServiceOrderamountQ12010] * 0 +
[ServiceOrderamountQ42010] * 0 +
[TransactioncountQ4MatCat1] * -0.08 +
[TransactioncountQ4MatCat2] * -0.08 +
[TransactioncountQ1MatCat5] * 0.35 +
[TransactioncountQ4MatCat5] * -0.31 +
[TransactioncountQ1STOCK] * 0.03 +
[CustomerCategoryB=NON-CATEGORISED] * 0.69 +
[CustomerCategoryC=L2 - DEALER] * -0.71 +
[CreditIndicator] * 0.27 +
[GeographicalBondIndicator] * 0.66 +
[Area=Area333] * 1.51 +
[Area=Area368] * 1.51 +
[Area=Area550] * 1.15 +
[Area=Area1502] * 1.51 +
[Area=Area1276] * 1.51 +
[Area=Area728] * -1.23 +
[Area=Area114] * 1.51 +
[Area=Area28] * 1.51 +
[Area=Area1440] * 1.51 +
[Area=Area685] * 1.51 +
[Area=Area1546] * 2.61 +
[Area=Area964] * 1.51 +
[Country] * -0.12

```

**Figure 4.22:** *logistic regression (SimpleLogistic) classifier results on dataset T*

SimpleLogistic is the Weka implementation of a linear Logistic Regression model based on the work of Sumner et al. (2005). It is an iterative algorithm which leads to attributes selection, in fact, in **Figure 4.22**, the coefficients are only assigned to a subset of explanatory variables and in some cases to values of the explanatory variables. For example, the Customer Category B variable has a coefficient only if the category B is of type NON-CATEGORISED, all other types are not relevant for this model.

The explanatory variables having a positive coefficient in the Class 1 vector increase the probability of being churner, moreover the larger is the weight, the stronger is the influence on the outcome. It is also worth considering the negative weights which tell what factors decrease such probability. To conclude, by looking at the weights, the inputs to consider are Area, Customer Category C, Customer Category B, Geographical Bond indicator, Transaction count Q1 MatCat5, Transaction count Q4 MatCat5, Credit Indicator.

Also for this model, to validate the results, the 10-fold cross validation technique has been adopted, giving the accuracy results highlighted in **Figure 4.23**. Correctly classified instances are 215 (75.17%) and incorrectly classified instances are 71 (24.83%).

```

Time taken to build model: 0.99 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      215           75.1748 %
Incorrectly Classified Instances    71           24.8252 %
Kappa statistic                    0.3221
Mean absolute error                 0.3218
Root mean squared error             0.4178
Relative absolute error             76.4048 %
Root relative squared error        91.1049 %
Total Number of Instances          286

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.92    0.64    0.77    0.92    0.838    0.783    N
                0.36    0.08    0.66    0.36    0.466    0.783    Y
Weighted Avg.   0.752    0.471    0.737    0.752    0.726    0.783

=== Confusion Matrix ===
  a  b  <-- classified as
184 16 | a = N
 55 31 | b = Y

```

**Figure 4.23:** SimpleLogistic classifier accuracy

To construct the Neural Network model, the MultilayerPerceptron algorithm has been selected from the Weka tool; this model uses the back propagation procedure described in the literature review. The results, consisting in a vector of weights for each node, are partially reported in **Figure 4.24** (only the first 3 nodes are reported and for each node only some weights shown).

```

Sigmoid Node 0
  Inputs  Weights
Threshold -11.13905763531526
Node 2    -0.7639599267985683
Node 3    -3.2181229117845898
[...]
Node 102   0.23612231774761008
Node 103   0.6187897612918618
Node 104   0.12976596488035785

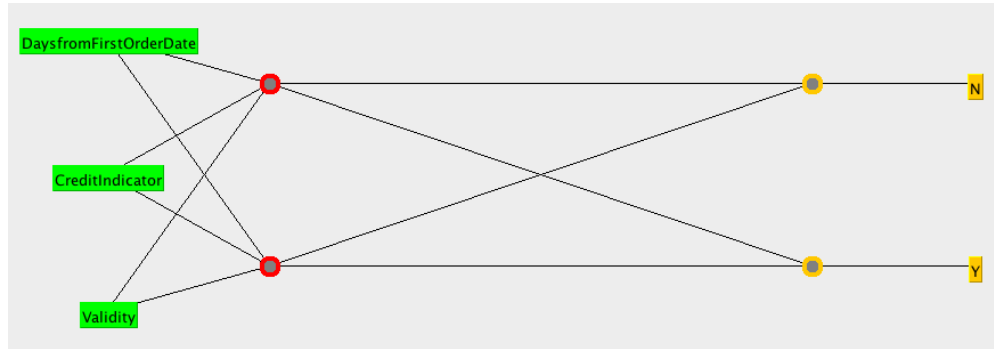
Sigmoid Node 1
  Inputs  Weights
Threshold 11.111151470315455
Node 2    0.701454291304606
Node 3    3.086797609173379
[...]
Node 102  -0.24103014001019063
Node 103  -0.5386211499664434
Node 104  -0.1210412625730859

Sigmoid Node 2
  Inputs  Weights
Threshold 0.04802526567696576
Attrib DaysfromFirstOrderDate -1.7154346498246897
Attrib PartsOrdercountQ12010 -0.23302034325903634
[...]
Attrib Area=Area964 -0.9516810737834336
Attrib Country      0.19210795204294803
Attrib Validity     -1.5996654533701573

```

**Figure 4.24:** neural network (MultilayerPerceptron) classifier results on dataset *T*

The network constructed consists in 104 neurons, the first two nodes (Node 0 and Node 1, see **Figure 4.24**) are the output nodes representing the Class 0 or 1 (N, not churner or Y, churner). All other 102 calculated nodes (Node 2, Node 3, ..., Node 102, Node 103, Node 104), which represents the hidden layer of the network, are connected to the two output nodes (for example Node 2 in **Figure 4.24**). It is evident that the inputs of the hidden layer nodes are instead the attributes of the dataset *T*. **Figure 4.25**, showing a small part of the network, clarifies the interconnection of the output (yellow), hidden (red) and input (green) nodes.



**Figure 4.25:** *partial representation of the Multilayer Perceptron network*

Unlike Logistic Regression and Decision Tree models, Neural Networks do not perform any attribute selection, as consequence all attributes give a contribution to the output.

Moreover, even looking at all the weights of each node, the model does not give any explanation on how and why certain results are produced, to put in another way, the model is opaque. As this is clearly a drawback of this algorithm, the non linear characteristic of the model is what makes the network so powerful (Berry and Linoff, 2004).

There have been attempts to extract rules from networks using the sensitivity analysis technique, which studies the relation between inputs and outputs. This is a purely empirical analysis and look at magnitude of variations in the output by changing one input at a time, clearly against the strict non linear interactions between input variables (Berry and Linoff, 2004).

On the other hand, since Logistic Regression can be considered a network consisting of only 1 node and due to its linear nature, it was, in that case, possible to compare the weights and select the contributors (relevant attributes).

**Figure 4.26** reports two important information: the accuracy of the model and the time taken to build it. The model correctly classifies 189 instances (66.08%) and incorrectly classifies 97 (33.92%) instances. Keeping in mind the small dataset *T*, it took over a minute (66.48 seconds) to build the network, considerably slower comparing the time taken by the previous two models, respectively 0.02 seconds and 0.99 seconds. It is worth taking into account this aspect for larger datasets.

```

Time taken to build model: 66.48 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      189           66.0839 %
Incorrectly Classified Instances    97           33.9161 %
Kappa statistic                    0.0877
Mean absolute error                 0.343
Root mean squared error             0.5502
Relative absolute error             81.4291 %
Root relative squared error        119.9717 %
Total Number of Instances          286

=== Detailed Accuracy By Class ===
                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.845    0.767    0.719    0.845    0.777      0.615    N
                0.233    0.155    0.392    0.233    0.292      0.615    Y
Weighted Avg.   0.661    0.583    0.621    0.661    0.631      0.615

=== Confusion Matrix ===
  a    b  <-- classified as
169  31   a = N
 66  20   b = Y

```

**Figure 4.26:** *MultilayerPerceptron classifier accuracy*

**Table 4.7** summarizes the accuracy of the three model implemented and discussed above.

Model	Algorithm	Correctly Classified	Incorrectly Classified
Decision Tree	J48	69.93%	30.07%
Logistic Regression	SimpleLogistic	75.17%	24.83%
Neural Network	MultilayerPerceptron	66.08%	33.92%

**Table 4.7:** *data mining models accuracy comparison*

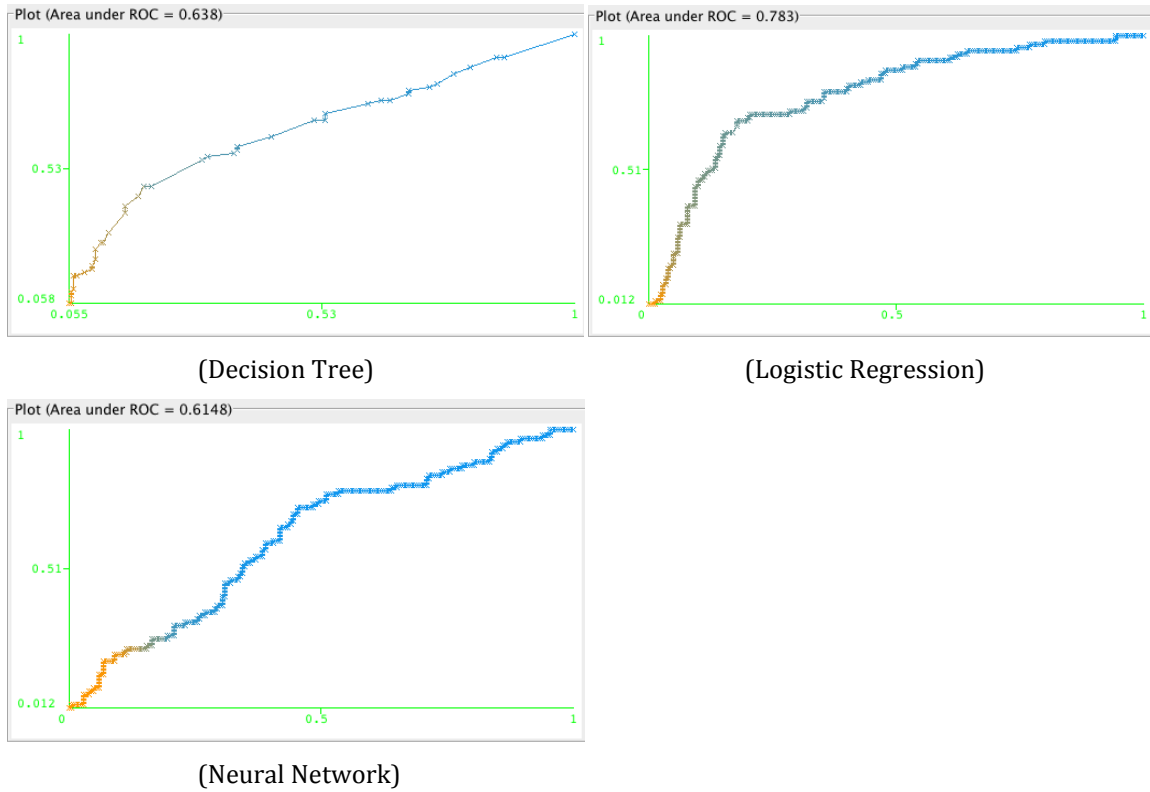
Accordingly to the accuracy, the better predictor is the Logistic Regression model, followed by the Decision Tree, which still has its value for the explicit rules that provides as opposed to the implicit rules of the logistic regression. The Neural Network is the less accurate and in general a more opaque model; its usage it is only justified if the accuracy is substantially greater than other models, which is not this case.

It is not enough deciding the best algorithm based only on the accuracy, also the errors committed need to be accounted (Vercellis, 2009). For this reason, the data mining tools such as Weka (see **Figure 4.19, 4.23, 4.26**) report for each model the so called confusion matrix. This matrix consists in 4 values: true positives, true negatives, false positives and false negatives. For the sake of the argument, consider the confusion matrix of the Neural Network of **Figure 4.26**. The true positives are 169, which are the correctly classified instance as Class 0 (N); the true negatives are 20, which are the correctly classified instance as Class 1 (Y); the false positives are 66, which are the instances classified as Class 0 (N) incorrectly, the false negatives are 31 which are the instances classified as Class 1 (Y) incorrectly. From this information two important indicators are calculated:

$$\text{True Positive Rate} = \text{True Positive} / (\text{True Positive} + \text{False Negative})$$

$$\text{False Positive Rate} = \text{False Positive} / (\text{False Positive} + \text{True Negative})$$

By plotting the True Positive Rate on the vertical axis against the False Positive Rate on the horizontal axis a ROC curve is obtained (Witten et al., 2011). ROC curve (receiver operating characteristic) is a popular technique to evaluate data mining model accuracy and compare different models: the better model is the one having greater area beneath the curve (Giudici, 2003; Vercellis, 2009; Witten et al., 2011).



**Figure 4.27:** ROC curve and ROC area

**Figure 4.27** shows the ROC curves generated by Weka for each implemented model. Once more, accordingly to the value of the area, the better model is the Logistic Regression (area = 0.783), followed by Decision Tree (area = 0.638) and Neural Network (0.6148). To understand the meaning of the ROC curves, say the false positive rate accepted is 20% (0.2 on the x axis), the equivalent true positive rate for Logistic Regression is around 70% (0.7 on the y axis), while Decision Tree is just below 50% and Neural Network less than 25%.

## 4.5 Evaluation and Deployment

The primary goal of the project, as defined in the Business Understanding stage, is to build a model with the best accuracy that is able to predict customer churners. This has been achieved in the Modelling stage with the Logistic Regression being the best predictor as result of the ROC curve analysis.

The secondary goal is to determine the variables that define customer churners. Also this goal has been accomplished in the Modelling stage but those variables need to be evaluated in the business context. Strictly speaking, the implicit rules generated by the Logistic Regression models require further analysis in order to understand how relevant they are in practice.

On one hand the aim of the project is to predict customers that are leaving, on the other hand to understand why they are leaving and take driven actions such as retention campaigns.

In general, the evaluation of the rules in the business context is a qualitative analysis and it is conducted presenting the results to the Client and summarizing the feedbacks. However, to establish if the results are successful or not in the business context, a basic quantitative approach is followed (derived from the Functional Point Analysis in Software projects).

For each variable under evaluation, a score of 0,1 or 2 is given. The score 0 indicates that the variable is not relevant in the business context and there is no evidence that may be used to retain customers; 1 means the variable has relevance but it is not easy to put in practice any action to retain customers (either financially not feasible or complex business processes to implement); 2 corresponds to those variables for which immediate actions can be taken to retain customers. If the total score is greater than the average results (assuming all variables score 1 on average), the results are considered successful in the business context, otherwise not. If unsuccessful, the predicting algorithm is still a valid result, but the implicit rules do not support or suggest any business action to stop the churners from leaving. As said in the Modelling stage, this is often the case when the best algorithm is the Neural Network for which the implicit rules are hard to interpret.

Appendix D reports the interview questions, the feedback and the score assigned by the Client. The questions are based on the explanatory variables resulting from the Logistic Regression, presented again here: Area, Customer Category C, Customer Category B, Geographical Bond indicator, Credit Indicator. Transaction count Q1 MatCat5 and Transaction count Q4 MatCat5 have been left out since they are numeric and not categorical, therefore hard to give a meaning for the same reasons given for Neural Networks.

The first question is on the Area. The list of areas sensible to churn (resulting from the Logistic Regression) seems to have no common denominator from a business perspective, however knowing the list a priori may be useful for the personnel. When placing an order, taking into account that the customer is sensible to churning, an ad hoc discount or promotion may be applied. This type of action cannot take place in the short term because the customer data, in particular the area information, requires quality checks and cleaning of the inconsistencies (problem already highlighted in Data Understanding stage). For those reasons the score assigned to Question #1 is 1.



Both questions #2 and #3, respectively on Customer Category C and Customer Category B, scored 2. The reason for this score is that, once it has been verified by the algorithm that churners mainly belongs to generic customer categories (applying standard end user price), a short term action can take place. The idea is to add new customer categories (of both type C and B) which satisfy individual needs of the customers. The system already allows to have customer category by material categories, therefore new category can be created having discounts on a set of material categories. To clarify this point, say a customer often buy filters and batteries belonging to the material category [P]. The Client can create a Customer Category B called GOLD – P which gives a 7% discount only on items from material category P.

Question #4 is on geographical bond, whereas local customers, accordingly to the algorithm results, have a 50% chance of being churners. From a strategic point of view the local customers are not treated in any different way to customers running their business in other locations. As such, the score given is 0: no business relevance and no action envisaged.

The final question is on Credit Indicator with the score of 1. The results from the Logistic Regression affirm that customers with credit facilities have more chances of churn compared to the ones paying by cash, which seems to be a contradiction. The explanation of this resides in the different level of credit facilities. There are customers with low credit limit and others with short day term. The former can use credit for small orders, the latter have a short credit period, usually within a week. Those two categories are comparable to pay by cash customers and most likely the churners indicated in the results. Usually the customers that find beneficial the credit facilities are the ones that are billed monthly and have high credit level. As possible long term action for this explanatory variable is to include in the data mining dataset *T* a more accurate definition of credit facilities, including day term and credit limit amount.

All the scores are summarized in **Table 4.8**.

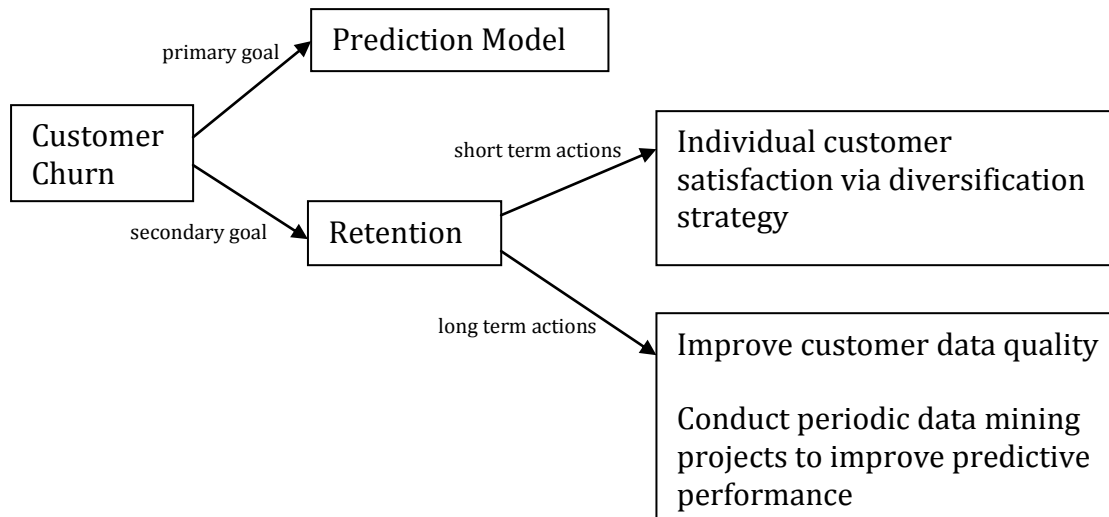
Question # - Variable	Score
Question #1 – Area	1
Question #2 – Customer Category C	2
Question #3 – Customer Category B	2
Question #4 – Geographical Bond Indicator	0
Question #5 – Credit Indicator	1
<b>Total</b>	6
<b>Average</b>	5

**Table 4.8:** *interview question scores, total and average*

The total score is 6, which is just above the average score 5 (assuming all questions are scored 1). The results can be considered moderately successful in the business context, this is in line with what discussed above, whereas the only 2 variables that can really be of use in the short term are Customer

Category C and B. Area and Credit Indicator instead have long term actions to take, while Geographical Bond has no business relevance.

The final stage of the project is the Deployment stage consisting in summarizing the knowledge acquired in a format that is understandable for Client's personnel, even though they have not been involved in the project. In other word a high level synthesis of the goals and outcomes.



**Figure 4.28:** data mining project goals and outcomes

As shown in **Figure 4.28**, on one hand, the primary goal is to determine what customers are going to leave, on the other hand, why those customers are leaving and what actions can be taken.

There are two categories of actions, short term and long term. The short term actions consist in creating new customer categories (in particular Customer Category C and B) in order to have a diversification strategy to support individual needs. The long term actions are more time consuming and require changes in the current business processes. First of all, the data gathered so far regarding customers require cleaning from noise and inconsistencies. Moreover, once the integrity of the data is achieved, the personnel need to be trained to enter accurate information and if this is not available an internal and consistent procedure must be agreed on how to proceed. Having more accurate and up to date data helps the performance of the predictive models. This means that the data mining project stages have to be repeated periodically. Another advantage of cyclical data mining project is that new questions arise from previous projects and answered in future ones and future ones benefit from the experience of the previous ones.

## 5 Conclusions

In this conclusive chapter the objectives are presented again in order to summarize how they have been met. What's more the contributions and limitations of the research are acknowledged and the next step proposed giving the researchers ideas on how to continue this work.

The dissertation concludes with some notes on ethic aspects, a delicate topic when analyzing data and treat differently individuals based on patterns found.

### 5.1 Contribution and Limitation

In the Introduction chapter the following objectives have been defined: to conduct an analysis of the Client customer, product and transactional data; to deliver a summary of the findings to support the Client retention plan; to identify, build and evaluate classification data mining models for predicting customer churn in the automotive industry; to identify which model is a better predictor in the Client case study and attempt to generalize the theory to the automotive industry.

The objective of analyzing the Client data has been met with the Data Understanding stage of the CRIP-DM model. The output of the stage is a set of charts showing aggregate information on customers, transactions, orders in the period under investigation. Moreover, as part of the Deployment stage, a high level summary of the mining project goals and outcomes suggests the Client stakeholders what short and long term actions should be taken in order to retain customers (retention plan objective).

The Modelling stage instead represents the contribution to the data mining field, stating that, among the classification models built, the Logistic Regression is potentially the best predictor in the context of the automotive industry. The exploratory variables defined in the Business Understanding and used in the Modelling stage are also another possible contribution as it suggests, by applying a framework, what information is required to build a data mining model in the automotive industry.

The results have been validated from a technical point of view using the 10-fold cross validation technique and the ROC curve analysis while the interview tested them in the business context.

However, the potential contributions to the data mining field required further work to be confirmed and being generalized to theory. The first limitation is that the mining models have been evaluated only on the Client case study: the results may be affected by specific business processes implemented by the Client and not necessarily true for other automotive dealerships. Another limitation of the project is that the three models have been built with standard parameters of the Weka tool: predictive performances may vary with different parameter values, more suitable for certain datasets.

### 5.2 Next Step

The results to be confirmed require additional data mining projects in a similar context. The focus of the next projects is to use similar sets of explanatory variables on several case studies and also compare the models not only with each other but also with the same model with different sets of parameter (not the standard ones).

Another area worth investigating is to expand the set of explanatory variables. The customers in this project have been considered individually and the variables show personal characteristics and trends

in their own activity. Accordingly to the work of Richter et al. (2010), the additional information to take into account when modelling retention is the social influence of social groups. The basic idea is that appealing deals from competitors or dissatisfaction with a service provided by the company investigated quickly circulates in social groups and it may affect churn decisions. Once again, the research conducted by Richter et al. (2010) is focused on mobile industry, therefore the challenge is to validate the results in the automotive industry.

Another interesting research area is calculating the customer potential value of churners and establish if this is greater than the cost of retention. By way of explanation, say 100 is the expected value from customers that are identified as churners and the cost of the retention campaign is estimated to be 120, the lost in the best case (all churners stay) is 20, therefore it is not worth retain them. Instead if the expected value of the churners is 150, the revenues are 30, therefore the retention plan is considered beneficial. In literature, the customer life time value (*LTV*) is defined as the total net income a company expect from a customer (Novo, 2004). The mathematical definition depends on many factors, one of the proposed models is the following (Rosset et al., 2002):

$$LTV = \int_0^{\infty} S(t)v(t)D(t)dt$$

whereas  $S(t)$  is the customer's churn probability,  $v(t)$  the customer's value over time,  $D(t)$  a discounting factor saying how much is worth today a future amount of money.

The  $S(t)$  probability function has been the outcome of this project, instead the estimation of  $v(t)$  is a possible further project, while  $D(t)$  can be set to 1 for simplicity.

Say  $C$  is the cost of the retention campaign, it is worth proceeding with it only if the following condition is true:

$$LTV - C > 0 \text{ otherwise } LTV > C$$

### 5.3 Data Mining and Ethics

The practice of data mining raises ethical and social issues as people generally do not know to what extent the stored data is used (Lawler and Molluzzo, 2005). In an open letter (Kim, 2003), the directors of ACM's SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) tried to clarify the difference between the research and the application of this technology.

On one side, the research aims at optimizing the models so that they produce ever more reliable results, therefore avoiding the danger of producing the so called false positive or false negative. For example, in categorizing customers based on their behaviours, it is clearly disfavour if a customer is assigned to a wrong category, for which he is not going to receive a certain benefit.

On the other side, the applications of those models are performed by specific companies on specific set of data for specific purposes. The main dangers to civil liberties are the misuse of the data collected, for example disseminating it without permission, or accessing in unauthorized ways. It is also fair to say that any information gathering technology (for instance database management, data warehousing, OLAP, biometric recognition), not just data mining, has this type of issues and, in general, the data analysts should follow ethical principles in order to respect the personal rights of individuals.

## References

- Ahn, J., Han, S. and Lee, Y. (2006) 'Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry', *Telecommunications Policy*, 10, pp 552- 568
- Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*, 3<sup>rd</sup> ed. Chichester: Wiley
- Bendapudi, N. and Berry, L. L. (2007) 'Customer's motivations for maintaining relationships with service providers', *Journal of Retailing*, Spring, 73 (1), pp 15-38
- Berne, C., Mugica, J. M. and Yague, M. J. (2001) 'The Effect of Variety Seeking on Customer Retention in Services', *Journal Of Retailing and Consumer Services*, 8, pp 335-345
- Berry, M. J. A. and Linoff, G. S. (2004) *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*, 2<sup>nd</sup> ed. Indianapolis: Wiley
- Bishop, C. M. (1995) *Neural networks for pattern recognition*, New York: Oxford University Press
- Brachman, R. and Anand, T. (1996) 'The Process of Knowledge Discovery in Databases: A Human Centered Approach', *AKDDM*, AAAI/MIT Press, pp 37-58
- Brachman, R. J. and Levesque, H. J. (1985) *Readings in knowledge representation*, San Francisco: Morgan Kaufmann
- Buckinx, W. and Van Den Poel, D. (2004) 'Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting', *European journal of operational research*, 164, pp 252-268
- Burez, J. and Van Den Poel, D. (2008) 'Separating financial from commercial customer churn: A modelling step towards resolving the conflict between the sales and credit department', *Expert Systems with Applications*, In Press, Corrected Proof
- Fayyad, U., Grinstein, G., Wierse, A. (2002) *Information visualization in data mining and knowledge discovery*, San Francisco: Morgan Kaufmann
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'Knowledge Discovery and Data Mining: Towards a Unifying Framework', *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, 1996, OR: AAAI Press, pp 82-88

- Giudici, P. (2003) *Applied Data Mining: Statistical Methods for Business and Industry*, Guildford, Surrey: Wiley
- Grinstead, C. and Snell, S. (1997) *Introduction to Probability*, Revised ed. American Mathematical Society. [Online] Available from:  
[http://www.dartmouth.edu/~chance/teaching\\_aids/books\\_articles/probability\\_book/amsbook.mac.pdf](http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf)  
[Accessed 9 October 2011]
- Hadden, J. (2008) *A Customer Profiling Methodology for Churn Prediction*, PhD thesis, Cranfield University
- Kim, H. and Yoon, C. (2004) 'Determinants of Subscriber Churn and Customer Loyalty in the Korean Mobile Telephony Market', *Telecommunications Policy*, 28, pp 751-765
- Kim, W. (2003) 'Data Mining Is NOT Against Civil Liberties', *Executive Committee on ACM Special Interest Group on Knowledge Discovery and Data Mining*, Austin, 28 July
- Lawler, J. and Molluzzo, J.C. (2005) 'A Study of Data Mining and Information Ethics in Information Systems Curricula', *Information Systems Education Conference ISECON Proceedings*, Columbus, OH, 2005
- Mallach E. (2000) *Decision Support and Data Warehouse Systems*, McGraw-Hill
- Mardia, K.C., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis (Probability and Mathematical Statistics)*, GB: Academic Press
- Mutanen, T. (2006) *Customer churn analysis – a case study*, VTT Business from technology Publications, (VTT-R-01184-06)
- Nordman, C. (2004) *Understanding Customer Loyalty and Disloyalty – The Effect of Loyalty-Supporting and –Repressing Factors*, Helsinki: Library Swedish School of Economics and Business Administration
- Novo, J. (2004) *Drilling Down: Turning Customer Data into Profits with a Spreadsheet*, U.S.: Booklocker Inc.
- Quinlan, J.R. (1986) 'Induction of decision trees', *Machine Learning*, 1(1), pp 81-106
- Quinlan, J.R. (1993) *C4.5: Programs for machine learning*, San Francisco: Morgan Kaufmann

Richter, Y., Yom-Tov, E. and Slonim, N. (2010) 'Predicting Customer Churn in Mobile Networks through Analysis of Social Groups', *Proceedings of the Tenth SIAM International Conference on Data Mining*, Columbus, Ohio, 2010, pp 732-741

Rosset, S., Neumann, E., Vatnik, U.E.N. and Idan, Y. (2002) 'Customer Lifetime Value Modeling and Its Use for Customer Retention Planning', *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Edmonton, Alberta, Canada, 23-26 July, New York: ACM

Rud, O.P. (2001) *Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management*, New York: Wiley

Shearer, C (2000) 'The CRISP-DM Model: The New Blueprint for Data Mining', *Journal of Data Warehousing*, 5 (4), pp 13-22

Stevens, S.S. (1946) 'On the theory of scales of measurement', *Science*, 103, pp 677-680

Sumner, M., Frank, E. and Hall, M. (2005): 'Speeding up Logistic Model Tree Induction', *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Porto, 2005, pp 675-683

Vercellis, C. (2009) *Business Intelligence: Data Mining and Optimization for Decision Making*, Milano: Wiley

Witten, I.H., Frank, E. and Hall, M.A. (2011) *Data mining: practical machine learning tools and techniques*, U.S.: Morgan Kaufmann

# Appendix A

## Project Definition

Project Title:

Customer Churn prediction for an Automotive Dealership using computational Data Mining.

Project Type:

Combined (client-based and academic research).

Problem To Solve:

The aim of the project is to identify explanatory variables in order to characterise the customers in the automotive industry and compare the different classification models for predicting customer churn.

Project Beneficiaries:

The project type (research strategy) is a combination of case study (client-based project) and academic research.

- Client: the dissertation, by using the client database as case study, provides a summary of the findings which can be evaluated and used in business decision makings.
- Data Mining area: identification of explanatory variables and prediction models comparison in automotive industry (generalization and theory testing).

Project Objectives:

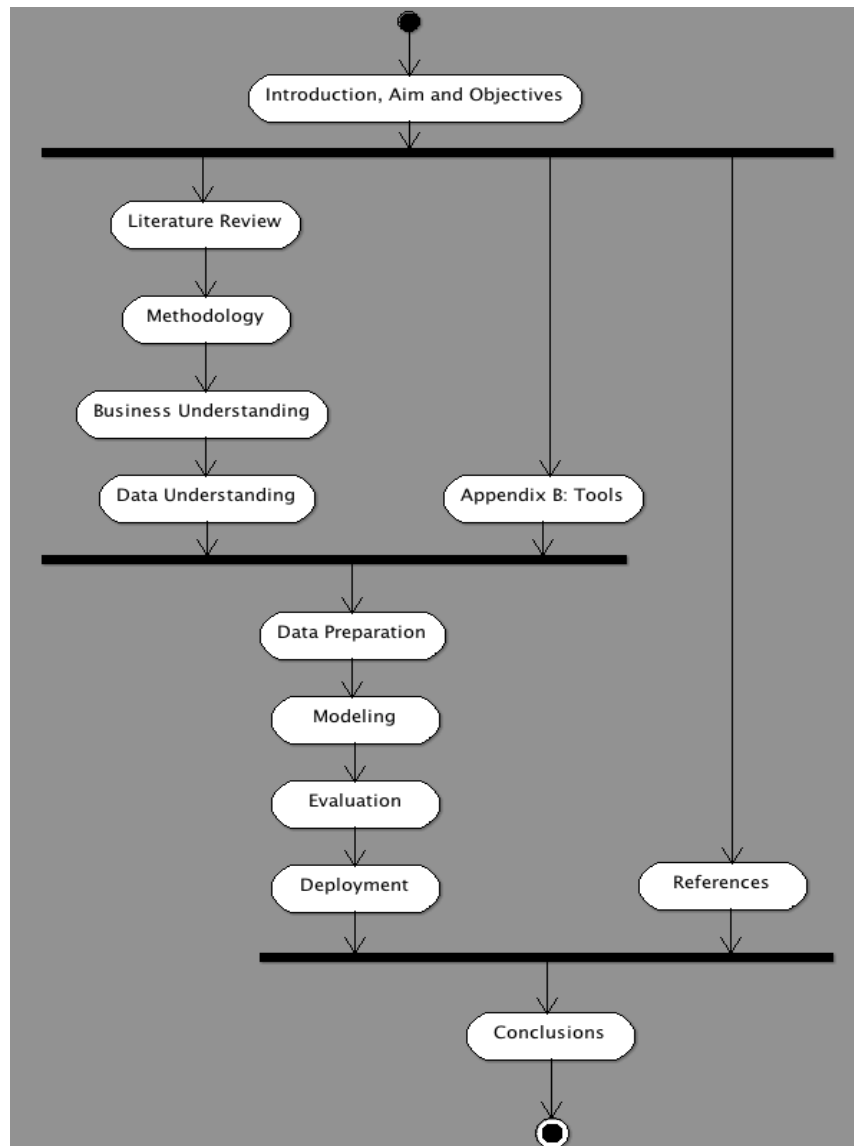
The aim of the project is to identify explanatory variables in order to characterise the customers in the automotive industry and compare the different classification models for predicting customer churn.

This can be achieved by meeting the following objectives:

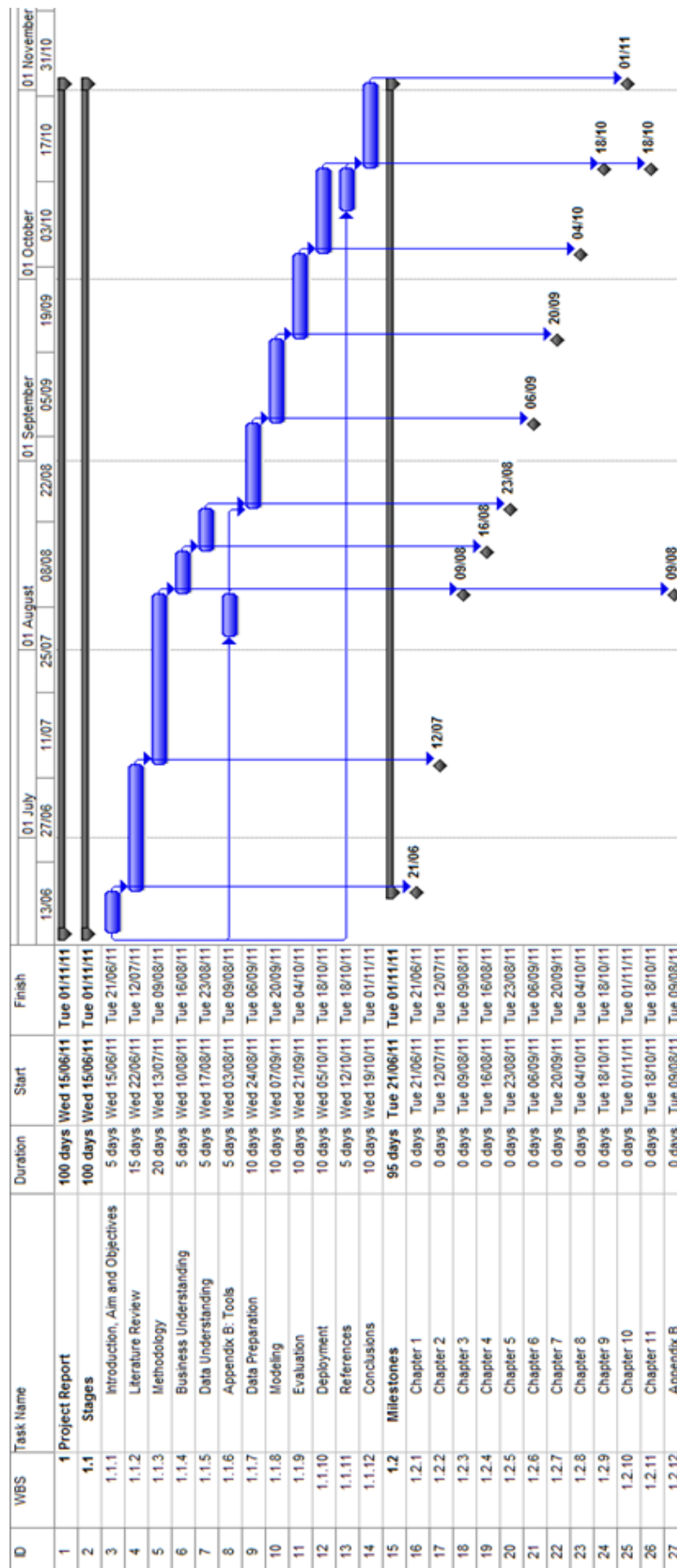
- To conduct an analysis of the Client customer, product and transactional data.
- To deliver a summary of the findings to support the Client retention plan.
- To identify, build and evaluate classification data mining models for predicting customer churn in the automotive industry.
- To identify which model is a better predictor of customer churn in the Client case study and attempt to generalize the theory to the automotive industry.



Proposed Project Structure:



## Proposed Project Plan:



## Appendix B

### Customer explanatory and target attributes

Attribute	Relevance
Days from First Order Date	Supporting-factor of type Dedication
Parts Order count Q1 2010	Supporting-factor of type Dedication
Parts Order count Q2 2010	Supporting-factor of type Dedication
Parts Order count Q3 2010	Supporting-factor of type Dedication
Parts Order count Q4 2010	Supporting-factor of type Dedication
Service Order count Q1 2010	Supporting-factor of type Dedication
Service Order count Q2 2010	Supporting-factor of type Dedication
Service Order count Q3 2010	Supporting-factor of type Dedication
Service Order count Q4 2010	Supporting-factor of type Dedication
Parts Order amount Q1 2010	Supporting-factor of type Dedication
Parts Order amount Q2 2010	Supporting-factor of type Dedication
Parts Order amount Q3 2010	Supporting-factor of type Dedication
Parts Order amount Q4 2010	Supporting-factor of type Dedication
Service Order amount Q1 2010	Supporting-factor of type Dedication
Service Order amount Q2 2010	Supporting-factor of type Dedication
Service Order amount Q3 2010	Supporting-factor of type Dedication
Service Order amount Q4 2010	Supporting-factor of type Dedication
Transaction count Q1 MatCat1	Supporting-factor of type Dedication
Transaction count Q2 MatCat1	Supporting-factor of type Dedication
Transaction count Q3 MatCat1	Supporting-factor of type Dedication
Transaction count Q4 MatCat1	Supporting-factor of type Dedication
Transaction count Q1 MatCat2	Supporting-factor of type Dedication
Transaction count Q2 MatCat2	Supporting-factor of type Dedication
Transaction count Q3 MatCat2	Supporting-factor of type Dedication
Transaction count Q4 MatCat2	Supporting-factor of type Dedication
Transaction count Q1 MatCat3	Supporting-factor of type Dedication
Transaction count Q2 MatCat3	Supporting-factor of type Dedication
Transaction count Q3 MatCat3	Supporting-factor of type Dedication
Transaction count Q4 MatCat3	Supporting-factor of type Dedication
Transaction count Q1 MatCat4	Supporting-factor of type Dedication
Transaction count Q2 MatCat4	Supporting-factor of type Dedication
Transaction count Q3 MatCat4	Supporting-factor of type Dedication

Transaction count Q4 MatCat4	Supporting-factor of type Dedication
Transaction count Q1 MatCat5	Supporting-factor of type Dedication
Transaction count Q2 MatCat5	Supporting-factor of type Dedication
Transaction count Q3 MatCat5	Supporting-factor of type Dedication
Transaction count Q4 MatCat5	Supporting-factor of type Dedication
Transaction count Q1 MatCatOthers	Supporting-factor of type Dedication
Transaction count Q2 MatCatOthers	Supporting-factor of type Dedication
Transaction count Q3 MatCatOthers	Supporting-factor of type Dedication
Transaction count Q4 MatCatOthers	Supporting-factor of type Dedication
Transaction count Q1 STOCK	Supporting-factor of type Dedication
Transaction count Q2 STOCK	Supporting-factor of type Dedication
Transaction count Q3 STOCK	Supporting-factor of type Dedication
Transaction count Q4 STOCK	Supporting-factor of type Dedication
Transaction count Q1 URGENT	Supporting-factor of type Dedication
Transaction count Q2 URGENT	Supporting-factor of type Dedication
Transaction count Q3 URGENT	Supporting-factor of type Dedication
Transaction count Q4 URGENT	Supporting-factor of type Dedication
Customer Category A	Supporting-factor of type Constraint
Customer Category B	Supporting-factor of type Constraint
Customer Category C	Supporting-factor of type Constraint
Credit Indicator	Supporting-factor of type Constraint
Lack of Alternative Indicator	Supporting-factor of type Constraint
Geographical Bond Indicator	Supporting-factor of type Constraint
Top Purchased Item price change 2010	Repressing-factor
Top Purchased Item alternative flag 2010	Repressing-factor
Customer category change indicator 2010	Repressing-factor
Average Service Completion days Q1	Repressing-factor
Average Service Completion days Q2	Repressing-factor
Average Service Completion days Q3	Repressing-factor
Average Service Completion days Q4	Repressing-factor
Area	Socio-demographic
Country	Socio-demographic
Validity	Socio-demographic
Churner indicator	Target Class

## Appendix C

### ORDER table

Attribute	Attribute type
Order ID	Categorical
Order Date	Categorical
Order Amount	Numerical
Order Currency	Categorical
Order Type	Categorical

### TRANSACTION table

Attribute	Attribute type
Transaction ID	Categorical
Material Category	Categorical
Delivery Priority	Categorical

### CUSTOMER table

Attribute	Attribute type
Customer ID	Categorical
Area ID	Categorical
Area Name	Categorical
Country ID	Categorical
Country Name	Categorical
Validity	Categorical

### **CUSTOMER RULE table**

<b>Attribute</b>	<b>Attribute type</b>
Customer ID	Categorical
Client Category A	Categorical
Client Category B	Categorical
Client Category C	Categorical
Version	Categorical
Credit	Categorical

### **ITEM table**

<b>Attribute</b>	<b>Attribute type</b>
Item ID	Categorical
Item Code	Categorical
Item Currency	Categorical
Item Price	Numerical
Price Start	Categorical
Price End	Categorical

### **ITEM ALTERNATIVE table**

<b>Attribute</b>	<b>Attribute type</b>
Item ID	Categorical
Item Code	Categorical
Item Relation Group	Categorical
Item Relation Type	Categorical

### **VEHICLE ORDER table**

<b>Attribute</b>	<b>Attribute type</b>
Job Order ID	Categorical
Order Date	Categorical
Completion Date	Categorical

## Appendix D

### Interview Introduction

The data mining algorithms produced a set of results to predict customer churners, in other words customers that are likely to leave. The aim of the interview is to find out if those results have a meaning in the business context so that driven actions can be taken such as retention campaigns. Note that the results are based on a subset of customers worth retaining (in a year they made purchases greater than a specific threshold).

For each question, the interviewee is asked to give some general comments and a score, defined as follow:

Score	Relevance	Possible Actions
0	Not relevant in business context	No action can be taken
1	Relevant in business context	No action can be taken (financially not feasible or complex business processes)
2	Relevant in business context	In the short terms, action can be taken

### Interview Question #1

Question #1 – Area	Score: 1
The following areas are sensible to churn: Area333, Area368, Area550, Area1502, Area1276, Area114, Area28, Area1440, Area685, Area1546 and Area964. Is there a common denominator between them? Knowing a priori that a customer belongs to one of those areas, can any action be taken to prevent churners?	
No evidence of a common denominator. Knowing the sensible areas a priori may be useful when placing an order and make additional ad hoc offers. However, the information stored for the customers (in particular the area) needs to be cleaned up (there are many duplicates) and from that moment data should be entered with diligence by the personnel.	

## Interview Question #1 - Notes

Question #1 - Area	Score: 4
<p>The following areas are sensible to churn: Area333, Area368, Area550, Area1502, Area1276, Area114, Area28, Area1440, Area685, Area1546 and Area964.</p> <p>Is there a common denominator between them?</p> <p>Knowing a priori that a customer belongs to one of those areas, can any action be taken to prevent churners?</p>	
<p>There is no evidence of a common denominator. Knowing the list of sensible area may be useful when placing orders: all the offers. The problem is data quality: customer info needs to be cleaned up. Data should be entered diligently by personnel.</p>	



## Interview Question #2

Question #2 – Customer Category C	Score: 2
<p>Customers with Customer Category C = L2 – DEALER (3-32% discount) have high probability of being loyal, while the majority of churners have Customer Category C = EUP. Also the categories such as L3 – DEALER (10%) and L4 – DEALER (12%) have a high number of loyal customers.</p> <p>Can the EUP customers be better differentiated giving a more specific category?</p>	
<p>Not all customers are dealers. Dealers are customers for high volume of purchases. However, today, for Customer Category C, the customers are divided only in two groups: end user and dealers. In the short term, new categories may be created based on individual needs. For instance, a customer may be interested in a particular set of materials, therefore a discount may be created for that material category but not the others (for which there is no interest).</p>	

## Interview Question #2 - Notes

## Question #2 - Customer Category C

Score:

2

Customers with Customer Category C = L2 - DEALER (3-32% discount) have high probability of being loyal, while the majority of churners have Customer Category C = EUP. Also the categories such as L3 - DEALER (10%) and L4 - DEALER (12%) have a high number of loyal customers.

Can the EUP customers be better differentiated giving a more specific category?

Not all customers are dealers.

Dealers make high volume purchases.

Today we have:

Customer Category C  $\begin{cases} \rightarrow \text{End User} \\ \rightarrow \text{Dealers} \end{cases}$

Other categories may be added in the short term, combining customer categories with material ones.

Natural category discounted

New Customer Category  
GOLD-P  
GOLD-Q  
GOLD-Y

	P	Q	Y	CS
GOLD-P	✓	x	x	x
GOLD-Q	x	✓	x	x
GOLD-Y	x	x	✓	x

A customer may be satisfied just buying category GOLD-P, if he/she buys items from material P.

### Interview Question #3

Question #3 - Customer Category B	Score: 2
Customer Category B = NON-CATEGORIZED contains the majority of the churners, while all other Customer Categories B have almost all loyal customers. As per Question #2, is there a way to have a better categorization?	
As per Question #2, in the short term also Customer Categories B needs to be reviewed an more specific categories may be created, combined to material categories to give individual discounts.	


### Interview Question #3 - Notes

Question #3 - Customer Category B	Score: (2)
Customer Category B = NON-CATEGORIZED contains the majority of the churners, while all other Customer Categories B have almost all loyal customers. As per Question #2, is there a way to have a better categorization?	
<p>Same consideration of question #2. Combine customer category with material category. This allows individual discounts.</p>	

## Interview Question #4

<b>Question #4 – Geographical Bond Indicator</b>	<b>Score: 0</b>
The results state that customers operating in the same area of the Client (geographical bond) are 50% churners and 50% loyal. However, it is worth noticing that there are no competitors in the same area. Any consideration on this?	
From a strategic point of view, there is no distinction between local area customers and customers from other areas. As such, this result is not relevant in the way the business is run.	


## Interview Question #4 – Notes

<b>Question #4 – Geographical Bond Indicator</b>	<b>Score:</b> 
The results state that customers operating in the same area of the Client (geographical bond) are 50% churners and 50% loyal. However, it is worth noticing that there are no competitors in the same area. Any consideration on this?	
<p>From a strategic point of view there is no distinction between local area customers and customers from other areas. This result is not relevant in the way the business is run.</p>	

## Interview Question #5

Question #5 – Credit Indicator	Score: 1
<p>Customers with credit facilities have more chances to be churners as opposed to customers paying by cash.</p> <p>Does this result sound controversial?</p>	
<p>Customers have different level of credit facilities.</p> <p>Some customers have low credit limit, which means can use credit for small orders, for big orders they still need to pay by cash the difference.</p> <p>Other customers have short days term, in other words the credit period is within a week and no more.</p> <p>Those two categories are comparable to pay by cash customers are a most likely the churners indicated in the results.</p> <p>Usually customers that benefits from credit facilities are the one that are billed monthly for all the purchases (30 days term) and have high credit level.</p> <p>In the long term, in a possible future second iteration of the mining project, a more accurate definition of credit facilities should be used and not just a simple indicator.</p> <p>Credit level and days term have to be considered.</p>	

## Interview Question #5 - Notes

Question #5 - Credit Indicator	Score: 
Customers with credit facilities have more chances to be churners as opposed to customers paying by cash.	
Does this result sound controversial?	
<p>Customers have different credit level and <del>stay</del> terms.</p> <p>Some customers have low credit limit, which means can use credit for small orders, for big orders still need to pay by cash.</p> <p>Other customers have short days than in other vendors like credit period is within a week and no more.</p> <p>Those 2 categories are not likely the churners indicated in the results.</p> <p>Usually customers having benefit from credit are the ones billed monthly (30 days term) and have high credit limit.</p> <p>In a possible next iteration of mining project or more accurate definition of credit facilities.</p>	

# Appendix E

## CD Content

### DataUnderstanding Folder

**SQL Folder:** sql queries and results to analyze the explanatory variables identified in the Business Understanding stage.

**Excel Folder:** charts generated using the results of the sql queries and included in the Project Report.

**Matlab Folder:** boxplot analysis to identify the customers to exclude from the project (so called outliers).

### DataPreparation Folder

**SQL Folder:** a shell script to execute the data preparation process, sql queries to construct the final derived table and the MySQL dump of the final table.

**Matlab Folder:** plot of the normal distribution with mean = 1333 and standard deviation = 825, used to define the Churner indicator (target class).

### Modelling Folder

**Weka:** results of the classification algorithms applied to the final table constructed in Data Preparation.

**DecisionTree Folder:** results and ROC area for evaluation.

**LogisticFunction Folder:** results and ROC area for evaluation.

**MultilayerPerceptron Folder:** results and ROC area for evaluation.

### Evaluation Folder

**Weka:** bivariate analysis (Churner indicator versus logistic regression explanatory variable) to prepare the interview questions.

**Interview:** questions and scan of the notes from the interview.

### Deployment Folder

Picture showing high level goals and outcomes of the projects.

## Appendix F

### Data Preparation SQL queries

```
set session group_concat_max_len=33553408;
```

#### #number and amount of parts orders and transactions

```
DROP TABLE IF EXISTS tmpPartsPurchasedDissertation;  
CREATE TABLE tmpPartsPurchasedDissertation (  
SELECT  
    idBuyer,  
    COUNT(DISTINCT IF(dtOrder >= var.pq1 AND dtOrder < var.pq2,idOrder,null)) AS ctPQ1,  
    COUNT(DISTINCT IF(dtOrder >= var.pq2 AND dtOrder < var.pq3,idOrder,null)) AS ctPQ2,  
    COUNT(DISTINCT IF(dtOrder >= var.pq3 AND dtOrder < var.pq4,idOrder,null)) AS ctPQ3,  
    COUNT(DISTINCT IF(dtOrder >= var.pq4 AND dtOrder < var.cq1,idOrder,null)) AS ctPQ4,  
    SUM(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2,IF(idCostCcyIso='EUR',(ctQtyMovement *  
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amPQ1,  
    SUM(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3,IF(idCostCcyIso='EUR',(ctQtyMovement *  
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amPQ2,  
    SUM(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4,IF(idCostCcyIso='EUR',(ctQtyMovement *  
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amPQ3,  
    SUM(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1,IF(idCostCcyIso='EUR',(ctQtyMovement *  
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amPQ4,  
    SUM(IF(dtOrder >= var.cq1 AND dtOrder < var.cq2,IF(idCostCcyIso='EUR',(ctQtyMovement *  
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amQ1,  
    COUNT(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2 and idCodeGrpCatMat="S",idOrder,null)) AS  
tSctPQ1,  
    COUNT(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3 and idCodeGrpCatMat="S",idOrder,null)) AS  
tSctPQ2,  
    COUNT(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4 and idCodeGrpCatMat="S",idOrder,null)) AS  
tSctPQ3,  
    COUNT(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1 and idCodeGrpCatMat="S",idOrder,null)) AS  
tSctPQ4,  
    COUNT(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2 and idCodeGrpCatMat="Q",idOrder,null)) AS  
tQctPQ1,  
    COUNT(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3 and idCodeGrpCatMat="Q",idOrder,null)) AS  
tQctPQ2,  
    COUNT(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4 and idCodeGrpCatMat="Q",idOrder,null)) AS  
tQctPQ3,  
    COUNT(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1 and idCodeGrpCatMat="Q",idOrder,null)) AS  
tQctPQ4,  
    COUNT(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2 and idCodeGrpCatMat="P",idOrder,null)) AS  
tPctPQ1,  
    COUNT(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3 and idCodeGrpCatMat="P",idOrder,null)) AS  
tPctPQ2,  
    COUNT(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4 and idCodeGrpCatMat="P",idOrder,null)) AS  
tPctPQ3,  
    COUNT(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1 and idCodeGrpCatMat="P",idOrder,null)) AS  
tPctPQ4,  
    COUNT(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2 and idCodeGrpCatMat="R",idOrder,null)) AS  
tRctPQ1,  
    COUNT(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3 and idCodeGrpCatMat="R",idOrder,null)) AS  
tRctPQ2,  
    COUNT(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4 and idCodeGrpCatMat="R",idOrder,null)) AS  
tRctPQ3,  
    COUNT(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1 and idCodeGrpCatMat="R",idOrder,null)) AS  
tRctPQ4,  
    COUNT(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2 and idCodeGrpCatMat="T",idOrder,null)) AS  
tTctPQ1,  
    COUNT(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3 and idCodeGrpCatMat="T",idOrder,null)) AS  
tTctPQ2,  
    COUNT(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4 and idCodeGrpCatMat="T",idOrder,null)) AS  
tTctPQ3,  
    COUNT(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1 and idCodeGrpCatMat="T",idOrder,null)) AS  
tTctPQ4,  
    COUNT(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2 and idCodeGrpCatMat not in  
("S","Q","P","R","T"),idOrder,null)) AS tOTctPQ1,
```

```

COUNT(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3 and idCodeGrpCatMat not in
("S","Q","P","R","T"),idOrder,null)) AS tOTctPQ2,
COUNT(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4 and idCodeGrpCatMat not in
("S","Q","P","R","T"),idOrder,null)) AS tOTctPQ3,
COUNT(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1 and idCodeGrpCatMat not in
("S","Q","P","R","T"),idOrder,null)) AS tOTctPQ4,
COUNT(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2 and idStatusPriorityDeliv=7,idOrder,null)) AS
tSTOCKctPQ1,
COUNT(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3 and idStatusPriorityDeliv=7,idOrder,null)) AS
tSTOCKctPQ2,
COUNT(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4 and idStatusPriorityDeliv=7,idOrder,null)) AS
tSTOCKctPQ3,
COUNT(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1 and idStatusPriorityDeliv=7,idOrder,null)) AS
tSTOCKctPQ4,
COUNT(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2 and idStatusPriorityDeliv<>7,idOrder,null)) AS
tURGENTctPQ1,
COUNT(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3 and idStatusPriorityDeliv<>7,idOrder,null)) AS
tURGENTctPQ2,
COUNT(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4 and idStatusPriorityDeliv<>7,idOrder,null)) AS
tURGENTctPQ3,
COUNT(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1 and idStatusPriorityDeliv<>7,idOrder,null)) AS
tURGENTctPQ4

```

FROM

```

mp_movementDenorm,
VAR_TCVS var,
mv_counterpartyFinancialToUpdate cpty,
mv_counterparty_view c

```

WHERE

```

idLocStockFrom IN (100,110,120) AND cancelled='N'
AND idTypeOrder IN (100,110) AND isJobOrder = 'N' AND idInvc is not null
AND cpty.idCpty=idBuyer
AND c.idCpty=cpty.idCpty AND c.idOwnCptyMaster=1
GROUP BY idBuyer
);

```

create index i1 on tmpPartsPurchasedDissertation(idBuyer);

#### #number and amount of service order

DROP TABLE IF EXISTS tmpServiceLabourPurchasedDissertation;

CREATE TABLE tmpServiceLabourPurchasedDissertation(

SELECT

```

idBuyer,
COUNT(DISTINCT IF(dtOrder >= var.pq1 AND dtOrder < var.pq2,idOrder,null)) AS ctPQ1,
COUNT(DISTINCT IF(dtOrder >= var.pq2 AND dtOrder < var.pq3,idOrder,null)) AS ctPQ2,
COUNT(DISTINCT IF(dtOrder >= var.pq3 AND dtOrder < var.pq4,idOrder,null)) AS ctPQ3,
COUNT(DISTINCT IF(dtOrder >= var.pq4 AND dtOrder < var.cq1,idOrder,null)) AS ctPQ4,
SUM(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2,IF(idCostCcyIso='EUR',(ctQtyMovement *
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amPQ1,
SUM(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3,IF(idCostCcyIso='EUR',(ctQtyMovement *
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amPQ2,
SUM(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4,IF(idCostCcyIso='EUR',(ctQtyMovement *
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amPQ3,
SUM(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1,IF(idCostCcyIso='EUR',(ctQtyMovement *
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amPQ4,
SUM(IF(dtOrder >= var.cq1 AND dtOrder < var.cq2,IF(idCostCcyIso='EUR',(ctQtyMovement *
amCost) * 1, ctQtyMovement * amCost * xrtEURActual),0)) AS amQ1,
AVG(IF(dtOrder >= var.pq1 AND dtOrder < var.pq2,DATEDIFF(dtJobOrdCompletion,dtOrder),0)) AS
qPQ1,
AVG(IF(dtOrder >= var.pq2 AND dtOrder < var.pq3,DATEDIFF(dtJobOrdCompletion,dtOrder),0))
AS qPQ2,
AVG(IF(dtOrder >= var.pq3 AND dtOrder < var.pq4,DATEDIFF(dtJobOrdCompletion,dtOrder),0))
AS qPQ3,
AVG(IF(dtOrder >= var.pq4 AND dtOrder < var.cq1,DATEDIFF(dtJobOrdCompletion,dtOrder),0))
AS qPQ4
FROM mp_serviceDenorm, VAR_TCVS var, mv_counterpartyFinancialToUpdate cpty,mv_counterparty_view
c
WHERE idInvc IS NOT NULL AND idTypeOrder=110 AND cpty.idCpty=idBuyer and cpty.idCpty=c.idCpty and
c.idOwnCptyMaster=1
GROUP BY idBuyer
);

```

create index i1 on tmpServiceLabourPurchasedDissertation(idBuyer);



**#get list of relevant customers (have at least a part or service order)**

```
drop table if exists tmpCounterpartylistDissertation;  
create table tmpCounterpartylistDissertation(idCpty int(11));  
create index i1 on tmpCounterpartylistDissertation(idCpty);  
insert into tmpCounterpartylistDissertation  
SELECT
```

```
    idBuyer as idCpty  
FROM  
    tmpPartsPurchasedDissertation  
  
UNION  
SELECT  
    idBuyer as idCpty  
FROM  
    tmpServiceLabourPurchasedDissertation
```

```
;
```

**#exclude outliers**

```
delete from tmpCounterpartylistDissertation where idCpty in  
(655,5457,20036,23695,26289,27892,29209,32826,70810);  
#extra check to make sure we only consider customers and not supplier  
delete cl from tmpCounterpartylistDissertation cl, mv_counterparty_view c where cl.idCpty = c.idCpty and  
c.idTypeUser <> 2;
```

**#define Area Numbers (for privacy)**

```
set @pos:=0;  
drop table if exists tmpAreaDissertation;  
create table tmpAreaDissertation(  
select  
    concat("Area",@pos:=@pos+1) as AreaNumber,  
    txTown,  
    count(*)  
from  
    mv_counterparty_view  
where  
    idTypeUser = 2 and idOwnCptyMaster = 1  
group by  
    txTown  
);  
create index i1 on tmpAreaDissertation(txCtown);
```

**#get customer categories**

```
drop table if exists tmpCustomerCategories;  
create table tmpCustomerCategories(  
select  
    c.idCpty, idGrpCpty, g.nmGrp, idGrpCpty2, g2.nmGrp as nmGrp2, idGrpCpty3, g3.nmGrp as  
nmGrp3, c.pmtCashOnly, area.AreaNumber, c.idCountry, m.idCodeFiscalKnown  
from  
    mv_counterparty_view c  
    JOIN tmpCounterpartylistDissertation cl ON cl.idCpty=c.idCpty  
    JOIN mv_counterpartyMaster m ON c.idCptyMaster = m.idCptyMaster and c.idOwnApp =  
m.idOwnApp  
    LEFT JOIN mp_group g ON c.idGrpCpty=g.idGrp and g.idOwnApp=1  
    LEFT JOIN mp_group g2 ON c.idGrpCpty2=g2.idGrp and g2.idOwnApp=1  
    LEFT JOIN mp_group g3 ON c.idGrpCpty3=g3.idGrp and g3.idOwnApp=1  
    LEFT JOIN tmpAreaDissertation area ON c.txCtown = area.txCtown  
where  
    c.idOwnCptyMaster = 1 and c.idTypeUser = 2  
);  
create index i1 on tmpCustomerCategories(idCpty);
```

**#get duration of relationship**

```
drop table if exists tmpRelationDurationDissertation;  
create table tmpRelationDurationDissertation(  
select idBuyer, min(dtOrder) as minDtOrder from view_orders_full where idMarket = "HU" and idTypeOrder  
in (100,110) and idOwnApp=1 and cancelled = "N" group by idBuyer  
);  
insert into tmpRelationDurationDissertation  
select idBuyer, min(dtOrder) as minDtOrder from view_orders_full_2008 where idMarket = "HU" and  
idTypeOrder in (100,110) and idOwnApp=1 and cancelled = "N" group by idBuyer;
```

```
insert into tmpRelationDurationDissertation
select idBuyer, min(dtOrder) as minDtOrder from view_orders_full_2007 where idMarket = "HU" and
idTypeOrder in (100,110) and idOwnApp=1 and cancelled = "N" group by idBuyer;
```

```
insert into tmpRelationDurationDissertation
select idBuyer, min(dtOrder) as minDtOrder from view_orders_full_2006 where idMarket = "HU" and
idTypeOrder in (100,110) and idOwnApp=1 and cancelled = "N" group by idBuyer;
```

```
insert into tmpRelationDurationDissertation
select idBuyer, min(dtOrder) as minDtOrder from view_orders_full_2005 where idMarket = "HU" and
idTypeOrder in (100,110) and idOwnApp=1 and cancelled = "N" group by idBuyer;
```

```
drop table if exists tmpRelationDuration2Dissertation;
create table tmpRelationDuration2Dissertation(
select idBuyer, min(minDtOrder) as minDtOrder from tmpRelationDurationDissertation group by idBuyer);
```

```
drop table if exists tmpRelationDurationFinalDissertation;
create table tmpRelationDurationFinalDissertation(
select idBuyer, datediff("2011-01-01",minDtOrder) as daysRelation from
tmpRelationDuration2Dissertation);
create index i1 on tmpRelationDurationFinalDissertation(idBuyer);
```

#### **#calculate geo bond and lack of alternatives flags**

```
drop table if exists tmpCptyAreaAnalysisDissertation;
create table tmpCptyAreaAnalysisDissertation(
select
    c.idCpty,
    if(AreaNumber in
("Area681","Area682","Area683","Area684","Area236","Area237","Area238","Area239","Area240","Area241",
"Area242","Area243","Area244","Area245","Area246","Area247","Area718","Area719","Area720","Area14",
05,"Area1406","Area1407","Area1408","Area1409","Area1410","Area1411","Area1412","Area715"), "N", "Y"
) as lackOfAlt,
    if(AreaNumber in ("Area177","Area178","Area179"), "Y", "N") as geoBond
from
    mv_counterparty_view c
    join tmpAreaDissertation area on c.txTown = area.txTown
    join tmpCounterpartylistDissertation l on c.idCpty=l.idCpty
group by
    c.idCpty
);
create index i1 on tmpCptyAreaAnalysisDissertation(idCpty);
```

#### **#excluding outliers, get top bought items by customers in 2010**

```
DROP TABLE IF EXISTS tmpTopItemPurchasedDissertation;
CREATE TABLE tmpTopItemPurchasedDissertation(
SELECT
    idBuyer,
    idCodeItemMAIN,
    count(*) as itemCount
FROM
    mp_movementDenorm,
    VAR_TCVS var,
    tmpCounterpartylistDissertation cpty,
    mv_counterparty_view c
WHERE
    year(dtOrder) = 2010 AND
    idLocStockFrom IN (100,110,120) AND cancelled='N'
    AND idTypeOrder IN (100,110) AND isJobOrder = 'N' AND idInvc is not null
    AND cpty.idCpty=idBuyer
    AND c.idCpty=cpty.idCpty AND c.idOwnCptyMaster=1
GROUP BY
    idBuyer, idCodeItemMAIN
ORDER BY
    idBuyer ASC, itemCount DESC
);
create index i1 on tmpTopItemPurchasedDissertation(idBuyer);
create index i2 on tmpTopItemPurchasedDissertation(idCodeItemMAIN);
```

#### **#get max prices for item in 2009 and 2010**

```
DROP TABLE IF EXISTS tmpItemPricesDissertation;
CREATE TABLE tmpItemPricesDissertation(
SELECT
```

```

        idCodeItemMAIN,
        max(IF(idCostCcyIso='EUR', amCost, amCost * xrtEURActual)*IF(year(dtOrder)<2010,1,0)) as
maxPrice2009,
        max(IF(idCostCcyIso='EUR', amCost, amCost * xrtEURActual)*IF(year(dtOrder)=2010,1,0)) as
maxPrice2010
FROM
    mp_movementDenorm,
    VAR_TCVS var,
    tmpCounterpartylistDissertation cpty,
    mv_counterparty_view c
WHERE
    idLocStockFrom IN (100,110,120) AND cancelled='N'
    AND idTypeOrder IN (100,110) AND isJobOrder = 'N' AND idInvc is not null
    AND cpty.idCpty=idBuyer
    AND c.idCpty=cpty.idCpty AND c.idOwnCptyMaster=1
GROUP BY
    idCodeItemMAIN
);
create index i1 on tmpItemPricesDissertation(idCodeItemMAIN);
DROP TABLE IF EXISTS tmpTopItemAndAltDissertation;
CREATE TABLE tmpTopItemAndAltDissertation(
select
    ic.idCodeItem as idCodeItemMAIN,
    ii.idInstanceItem,
    count(*) as relCount
from
    mp_itemRel ir,
    mp_itemRelContent irc,
    mp_itemInstance ii,
    mp_itemCode ic
where
    ir.idRel = irc.idRel and ir.idTypeRel = "ALT"
    and irc.idInstanceItem = ii.idInstanceItem
    and ii.idOwnApp=1 and ii.idItem = ic.idItem
    and ic.idOwnApp=ii.idOwnApp
    and ic.idTypeCodeItem=10 and ic.idOwnCptyMaster=1
group by
    ic.idCodeItem
);
create index i1 on tmpTopItemAndAltDissertation(idCodeItemMAIN);
set @num := 0; set @idBuyer := "";

```

**#select top item for each customer in 2010 and check if the price is increased compared to 2009 and if there are alternatives**

```

DROP TABLE IF EXISTS tmpTopItemAndPriceAndAltDissertation;
CREATE TABLE tmpTopItemAndPriceAndAltDissertation(
select
    item.idBuyer,
    item.idCodeItemMAIN,
    item.itemCount,
    itemP.maxPrice2010 - itemP.maxPrice2009 as priceDiff,
    if(alt.idCodeItemMAIN is not null,"Y","N") as isThereAlt,
    @num := if(@idBuyer = item.idBuyer, @num+1,1) as subCount,
    @idBuyer:=item.idBuyer as idBuyerT
from
    (tmpTopItemPurchasedDissertation item,
    tmpItemPricesDissertation itemP)
LEFT JOIN tmpTopItemAndAltDissertation alt ON item.idCodeItemMAIN=alt.idCodeItemMAIN
where
    item.idCodeItemMAIN = itemP.idCodeItemMAIN
    and itemP.maxPrice2010 > 0 and itemP.maxPrice2009 > 0
having
    subCount = 1 and
    priceDiff > 0
);
create index i1 on tmpTopItemAndPriceAndAltDissertation(idBuyer);

```

**#excluding outliers, get customer whose categories changed in 2010**

```

DROP TABLE IF EXISTS tmpCustomerCategoryChangeDissertation;
CREATE TABLE tmpCustomerCategoryChangeDissertation(
SELECT
    cpty.idCpty,

```

```

        c.idCptyEntity,
        count(*) as changes
FROM
    tmpCounterpartylistDissertation cpty,
    mv_counterparty_view c,
    mv_counterpartyClientRule cr
WHERE
    c.idCpty = cpty.idCpty
    and c.idCptyEntity = cr.idCptyEntity
    and cr.idOwnCptyMaster = 1
    and year(cr.tmUpdate) = 2010
GROUP BY
    cpty.idCpty
);
create index i1 on tmpCustomerCategoryChangeDissertation(idCpty);

#build final dataset
#count to make sure the joins work ok
SELECT
    count(*)
FROM
    tmpCounterpartylistDissertation cpty
    JOIN tmpRelationDurationFinalDissertation rel ON cpty.idCpty = rel.idBuyer
    JOIN tmpCustomerCategories cat ON cpty.idCpty = cat.idCpty
    JOIN tmpCptyAreaAnalysisDissertation area ON cpty.idCpty = area.idCpty
    LEFT JOIN tmpPartsPurchasedDissertation parts ON cpty.idCpty = parts.idBuyer
    LEFT JOIN tmpServiceLabourPurchasedDissertation service ON cpty.idCpty = service.idBuyer
    LEFT JOIN tmpTopItemAndPriceAndAltDissertation topitem ON cpty.idCpty = topitem.idBuyer
    LEFT JOIN tmpCustomerCategoryChangeDissertation catchange ON cpty.idCpty = catchange.idCpty
;

```

```

#final select
drop table if exists tmpDATASETDissertation;
create table tmpDATASETDissertation(
SELECT
    concat("CUSTOMER",cpty.idCpty) as customerID,
    rel.daysRelation as DaysfromFirstOrderDate,
    ifnull(parts.ctPQ1,0) as PartsOrdercountQ12010,
    ifnull(parts.ctPQ2,0) as PartsOrdercountQ22010,
    ifnull(parts.ctPQ3,0) as PartsOrdercountQ32010,
    ifnull(parts.ctPQ4,0) as PartsOrdercountQ42010,
    ifnull(service.ctPQ1,0) as ServiceOrdercountQ12010,
    ifnull(service.ctPQ2,0) as ServiceOrdercountQ22010,
    ifnull(service.ctPQ3,0) as ServiceOrdercountQ32010,
    ifnull(service.ctPQ4,0) as ServiceOrdercountQ42010,
    ifnull(parts.amPQ1,0) as PartsOrderamountQ12010,
    ifnull(parts.amPQ2,0) as PartsOrderamountQ22010,
    ifnull(parts.amPQ3,0) as PartsOrderamountQ32010,
    ifnull(parts.amPQ4,0) as PartsOrderamountQ42010,
    ifnull(parts.amQ1,0) as PartsOrderamountQ12011,
    ifnull(service.amPQ1,0) as ServiceOrderamountQ12010,
    ifnull(service.amPQ2,0) as ServiceOrderamountQ22010,
    ifnull(service.amPQ3,0) as ServiceOrderamountQ32010,
    ifnull(service.amPQ4,0) as ServiceOrderamountQ42010,
    ifnull(service.amQ1,0) as ServiceOrderamountQ12011,
    ifnull(parts.tSctPQ1,0) as TransactioncountQ1MatCat1,
    ifnull(parts.tSctPQ2,0) as TransactioncountQ2MatCat1,
    ifnull(parts.tSctPQ3,0) as TransactioncountQ3MatCat1,
    ifnull(parts.tSctPQ4,0) as TransactioncountQ4MatCat1,
    ifnull(parts.tQctPQ1,0) as TransactioncountQ1MatCat2,
    ifnull(parts.tQctPQ2,0) as TransactioncountQ2MatCat2,
    ifnull(parts.tQctPQ3,0) as TransactioncountQ3MatCat2,
    ifnull(parts.tQctPQ4,0) as TransactioncountQ4MatCat2,
    ifnull(parts.tPctPQ1,0) as TransactioncountQ1MatCat3,
    ifnull(parts.tPctPQ2,0) as TransactioncountQ2MatCat3,
    ifnull(parts.tPctPQ3,0) as TransactioncountQ3MatCat3,
    ifnull(parts.tPctPQ4,0) as TransactioncountQ4MatCat3,
    ifnull(parts.tRctPQ1,0) as TransactioncountQ1MatCat4,
    ifnull(parts.tRctPQ2,0) as TransactioncountQ2MatCat4,
    ifnull(parts.tRctPQ3,0) as TransactioncountQ3MatCat4,
    ifnull(parts.tRctPQ4,0) as TransactioncountQ4MatCat4,
    ifnull(parts.tTctPQ1,0) as TransactioncountQ1MatCat5,

```

```

ifnull(parts.tTctPQ2,0) as TransactioncountQ2MatCat5,
ifnull(parts.tTctPQ3,0) as TransactioncountQ3MatCat5,
ifnull(parts.tTctPQ4,0) as TransactioncountQ4MatCat5,
ifnull(parts.tOTctPQ1,0) as TransactioncountQ1MatCatOthers,
ifnull(parts.tOTctPQ2,0) as TransactioncountQ2MatCatOthers,
ifnull(parts.tOTctPQ3,0) as TransactioncountQ3MatCatOthers,
ifnull(parts.tOTctPQ4,0) as TransactioncountQ4MatCatOthers,
ifnull(parts.tSTOCKctPQ1,0) as TransactioncountQ1STOCK,
ifnull(parts.tSTOCKctPQ2,0) as TransactioncountQ2STOCK,
ifnull(parts.tSTOCKctPQ3,0) as TransactioncountQ3STOCK,
ifnull(parts.tSTOCKctPQ4,0) as TransactioncountQ4STOCK,
ifnull(parts.tURGENTctPQ1,0) as TransactioncountQ1URGENT,
ifnull(parts.tURGENTctPQ2,0) as TransactioncountQ2URGENT,
ifnull(parts.tURGENTctPQ3,0) as TransactioncountQ3URGENT,
ifnull(parts.tURGENTctPQ4,0) as TransactioncountQ4URGENT,
cat.nmGrp as CustomerCategorA,
cat.nmGrp2 as CustomerCategoryB,
cat.nmGrp3 as CustomerCategoryC,
cat.pmtCashOnly as CreditIndicator,
area.lackOfAlt as LackofAlternativeIndicator,
area.geoBond as GeographicalBondIndicator,
if(topitem.idBuyer is not null,"Y","N") as TopPurchasedItempricechange2010,
ifnull(topitem.isThereAlt,"N") as TopPurchasedItemalternativeflag2010,
if(catchchange.idCpty is not null,"Y","N") as Customercategorychangeindicator2010,
ifnull(service.qPQ1,0) as AverageServiceCompletiondaysQ1,
ifnull(service.qPQ2,0) as AverageServiceCompletiondaysQ2,
ifnull(service.qPQ3,0) as AverageServiceCompletiondaysQ3,
ifnull(service.qPQ4,0) as AverageServiceCompletiondaysQ4,
cat.AreaNumber as Area,
cat.idCountry as Country,
cat.idCodeFiscalKnown as Validity,
"N" as Churnerindicator
FROM
tmpCounterpartylistDissertation cpty
JOIN tmpRelationDurationFinalDissertation rel ON cpty.idCpty = rel.idBuyer
JOIN tmpCustomerCategories cat ON cpty.idCpty = cat.idCpty
JOIN tmpCptyAreaAnalysisDissertation area ON cpty.idCpty = area.idCpty
LEFT JOIN tmpPartsPurchasedDissertation parts ON cpty.idCpty = parts.idBuyer
LEFT JOIN tmpServiceLabourPurchasedDissertation service ON cpty.idCpty = service.idBuyer
LEFT JOIN tmpTopItemAndPriceAndAltDissertation topitem ON cpty.idCpty = topitem.idBuyer
LEFT JOIN tmpCustomerCategoryChangeDissertation catchchange ON cpty.idCpty = catchchange.idCpty
);

#filter the dataset by valuable customers only
DROP TABLE IF EXISTS tmpDataSetFilterDissertation;
CREATE TABLE tmpDataSetFilterDissertation(
select
customerID,

sum(PartsOrderamountQ12010+PartsOrderamountQ22010+PartsOrderamountQ32010+PartsOrderamountQ
42010+ServiceOrderamountQ12010+ServiceOrderamountQ22010+ServiceOrderamountQ32010+ServiceOr
deramountQ42010) as totalAmount2010,
sum(PartsOrderamountQ12011+ServiceOrderamountQ12011) as q12011,
sum(PartsOrderamountQ12010+ServiceOrderamountQ12010) as q12010
from
tmpDATASETDissertation
group by
customerID
having
totalAmount2010 > 500
);
CREATE INDEX i1 on tmpDataSetFilterDissertation(customerID);

#filter final dataset removing not valuable customers
DELETE t FROM tmpDATASETDissertation t LEFT JOIN tmpDataSetFilterDissertation f ON t.customerID =
f.customerID WHERE f.customerID IS NULL;
SELECT COUNT(*) FROM tmpDATASETDissertation;

#define Churner indicator (target class)
UPDATE tmpDATASETDissertation t, tmpDataSetFilterDissertation f SET t.Churnerindicator = "Y" WHERE
t.customerID = f.customerID AND q12010 > 0 and q12011 = 0;

```